Prédiction en grande dimension, Extrêmes et Réseaux

Charles-Elie Rabier

ISEM, Institut des Sciences de l'Evolution de Montpellier
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique

Première partie :

A propos de la prédiction en grande dimension

A propos de la prédiction en grande dimension

avec Brigitte Mangin

LIPM, Laboratoire des Interactions Plantes et Microorganismes, Toulouse

et Simona Grusea

Institut National des Sciences Appliquées de Toulouse
Institut de Mathématiques de Toulouse

Plan

Plan

Sélection génomique = statistique en grande dimension + apprentissage

Training population

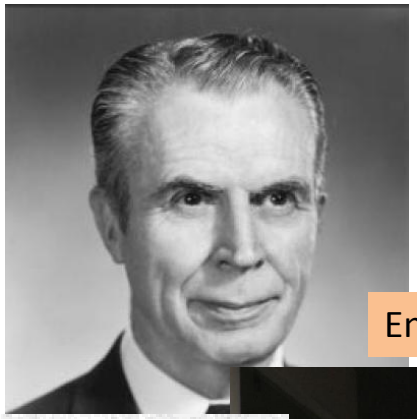
$$Y = \mu + \underbrace{f(M, \theta)}_{g} + \varepsilon$$ ← Model

$$\hat{\theta}$$ ← Estimation method

Test population (selection candidates)

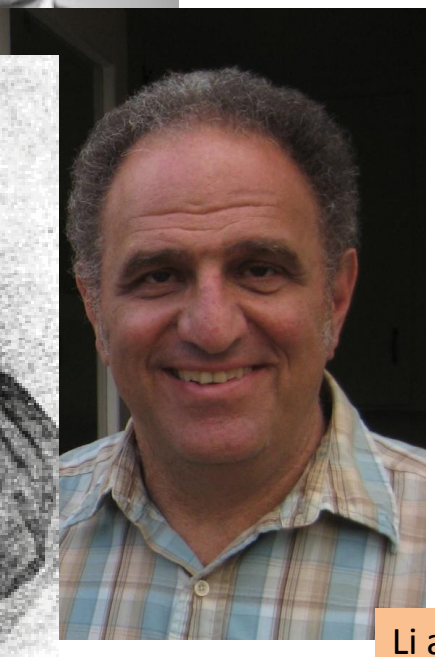
$$\underbrace{f(M, \hat{\theta})}_{\hat{g}}$$ ← Prediction

Cadres statistiques



C.R. Henderson
Modèle mixte

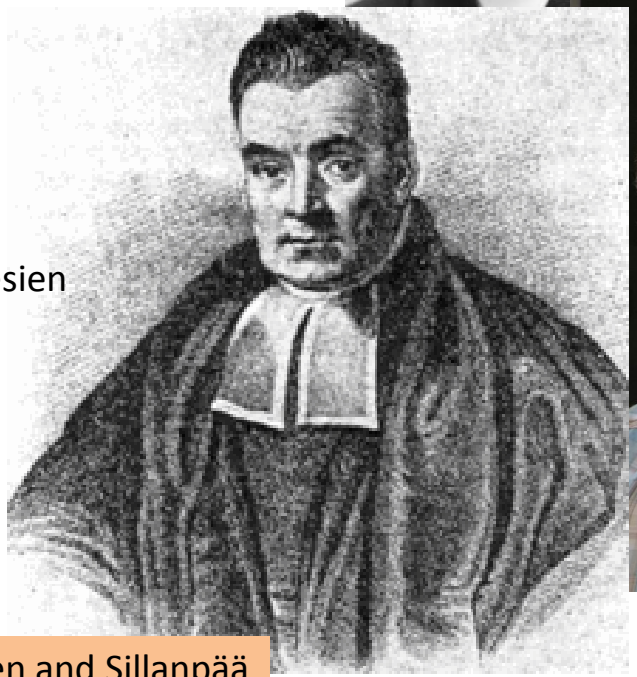
Endelman



R. Tibshirabi
Régression pénalisée

Li and Sillanpää

T. Bayes
Modèle bayésien



Kärkkäinen and Sillanpää

Classement

- En général, le classement des méthodes est Bayes $\geq$ Régressions pénalisées $>$ Modèle mixte
- Les méthodes ont moins d'influence que le nombre de variables explicatives, la taille de l' échantillon d'apprentissage, le signal
- Les distributions de probabilité des échantillons d'apprentissage et de validation (TEST) jouent également un rôle important

Cet exposé : focus sur la régression Ridge (Pénalité L2)

Statistique en grande dimension

Objectif : Prédire une variable continue

à l'aide d'un grand nombre de régresseurs

Modèle causal* (Q vrais régresseurs)

Echantillon d'apprentissage de taille n,
$\theta^\star$ vecteur d'effets, $M^\star$ matrice de mesures,

$$Y = M^\star \theta^\star + e$$

où $Y = (Y_1, ..., Y_n)'$, $\theta^\star = (\theta_1^\star, ..., \theta_Q^\star)'$, $e \sim N(0, \sigma_e^2 I_n)$

Modèle Bayésien de prédiction (K régresseurs, où $K >> n$)

θ vecteur d'effets, M matrice de mesures

$$Y = M\theta + \varepsilon$$

où $Y = (Y_1, ..., Y_n)'$, $\theta = (\theta_1, ..., \theta_K)' \sim N(0, \sigma_\theta^2 I_K)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$, $\varepsilon_j \perp\!\!\!\perp \theta_k$

On supposera que le modèle de prédiction contient les vrais régresseurs ...
Autrement dit, chaque colonne de $M^\star$ est une colonne de M

Phase d'apprentissage

Loi jointe de θ et Y

$$\begin{pmatrix} \theta \\ Y \end{pmatrix}_{|M} \sim N\left(0, \begin{pmatrix} \sigma_\theta^2 I_K & \sigma_\theta^2 M' \\ \sigma_\theta^2 M & \sigma_\theta^2 M M' + \sigma_\varepsilon^2 I_n \end{pmatrix}\right)$$

Estimateur $\hat{\theta}$ de θ

$$\begin{aligned} \hat{\theta} &= \mathbb{E}(\theta \mid Y) = M' \left(MM' + \lambda I_n\right)^{-1} Y \quad \text{où} \quad \lambda = \sigma_\varepsilon^2/\sigma_\theta^2 \\ &= \left(M'M + \lambda I_K\right)^{-1} M' Y \end{aligned}$$

i.e. Régression Ridge (Pénalité L2) avec paramètre $\lambda = \sigma_\varepsilon^2/\sigma_\theta^2$

$$\hat{\theta} = \text{argmin}_\theta \quad \|Y - M\theta\|^2 + \lambda \|\theta\|^2$$

Echantillon de validation + critère d'accuracy

- Soit un individu TEST noté new

$$Y_{\text{new}} = m_{\text{new}}^{\star\prime}\,\theta^{\star} + e_{\text{new}} \quad \text{où} \quad e_{\text{new}} \sim N(0, \sigma_e^2)$$
$$\text{et} \quad m_{\text{new}}^{\star} \text{ vecteur de mesures de l'individu new}$$

- Prédiction de la variable continue Y_{new}

$$\hat{Y}_{\text{new}} = m_{\text{new}}'\hat{\theta} = m_{\text{new}}' M' \left(MM' + \lambda I_n\right)^{-1} Y$$
$$= m_{\text{new}}' \left(M'M + \lambda I_K\right)^{-1} M' Y$$

$\Rightarrow$ Critère d'accuracy (i.e. précision de la prédiction)

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{\text{new}}\right)\text{Var}\left(Y_{\text{new}}\right)}} \quad \text{avec } m_{\text{new}} \text{ et } m_{\text{new}}^{\star} \text{ aléatoires, } M \text{ fixe}$$

A propos de l'aléatoire dans notre analyse

Echantillon d'apprentissage :

- l'analyse est conditionnelle à M et $M^\star$
- le vecteur $Y = (Y_1, ..., Y_n)'$ reste aléatoire car le bruit e est aléatoire
- $\hat{\theta} = M' \left(MM' + \lambda I_n\right)^{-1} Y$ est aléatoire

Echantillon de validation :

- m_{new}, $m^\star_{\text{new}}$ et Y_{new} sont aléatoires

Mes études portant sur l'accuracy

Critère d'accuracy (i.e. précision de la prédiction)

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{\text{new}}\right)\text{Var}\left(Y_{\text{new}}\right)}}$$

- Etude appliquée : R., Barre ... Mangin (Plos One, 2016)
  collaborations avec Biologistes (SupAgro, INRA, Danemark)
- Etude théorique :
  R., Mangin, Grusea (Scandinavian Journal of Statistics, 2018)
  convergence (en fonction des val singulières, proj signal, λ)
  + autre estimateur basé sur un espace de dimension plus faible

Plan

Résultat sur l'accuracy (i.e. précision de la prédiction)

Comme le modèle de prédiction contient les vrais régresseurs, on notera par abus de notation

- $\theta^\star$ vecteur sparse de dimension K

alors, le modèle causal se réécrit sous la forme

$$Y = M\theta^\star + e \quad \text{où} \quad Y = (Y_1, ..., Y_n)' \ , \ e \sim N(0, \sigma_e^2 I_n).$$

Théorème (Rabier Barre ... Mangin, Plos One 2016)

Supposons que M est connu, et que e, m_{new} et e_{new} sont aléatoires, alors

$$\rho = \frac{\theta^{\star\prime} \ Var(m_{\text{new}}) \, M' V^{-1} M\theta^\star}{\left\{\sigma_e^2 \mathbb{E}\left(\|m'_{\text{new}} M' V^{-1}\|^2\right) + \theta^{\star\prime} M' V^{-1} M Var(m_{\text{new}}) \, M' V^{-1} M\theta^\star\right\}^{1/2} \Omega^{1/2}}$$

où $V = MM' + \lambda I_n$ et $\Omega = \text{Var}\,(m'_{\text{new}}\theta^\star) + \sigma_e^2$

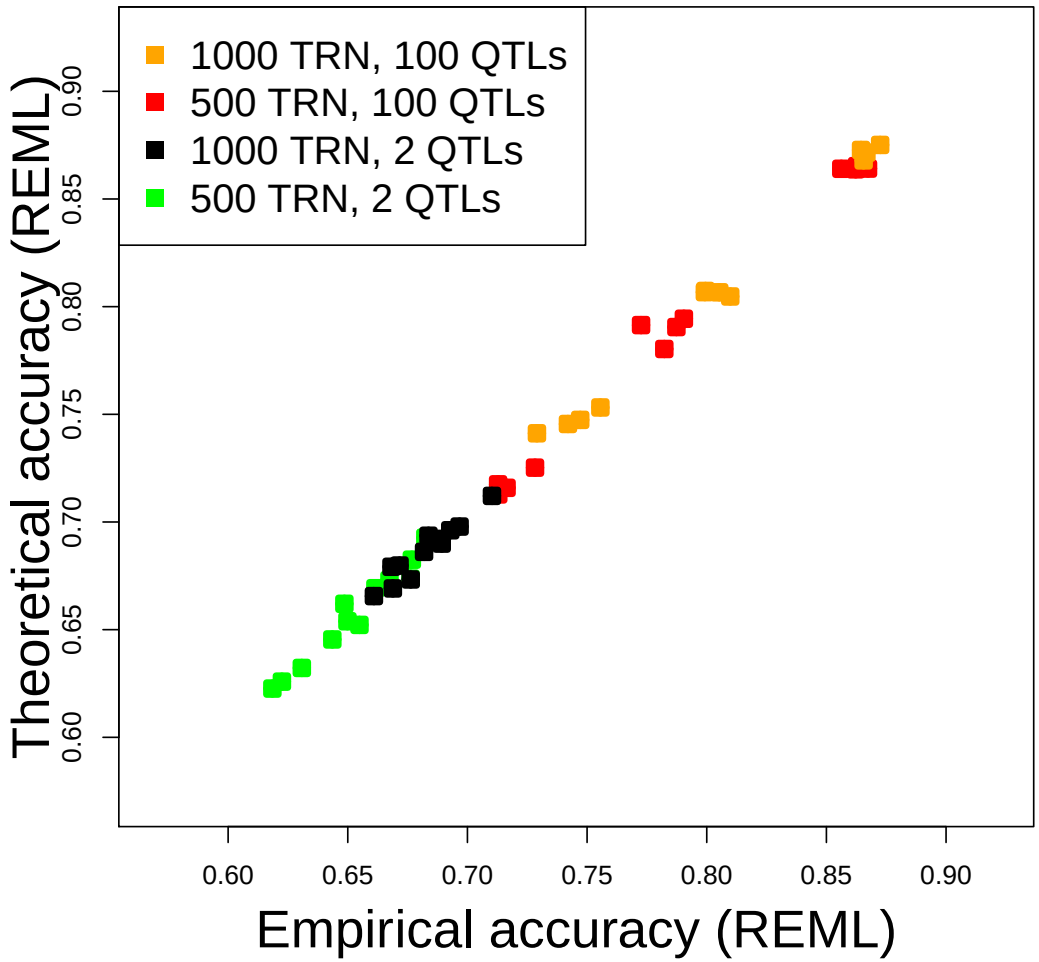
Au dénominateur, $\mathbb{E}\left\{\text{Var}\left(\hat{Y}_{\text{new}} \mid m_{\text{new}}\right)\right\} = \sigma_e^2 \mathbb{E}\left(\|m'_{\text{new}} M' V^{-1}\|^2\right)$

et $\text{Var}\left\{\mathbb{E}\left(\hat{Y}_{\text{new}} \mid m_{\text{new}}\right)\right\} = \theta^{\star\prime} M' V^{-1} M \text{Var}\,(m_{\text{new}}) \, M' V^{-1} M\theta^\star$

Vérification de la formule sur données simulées en génomique

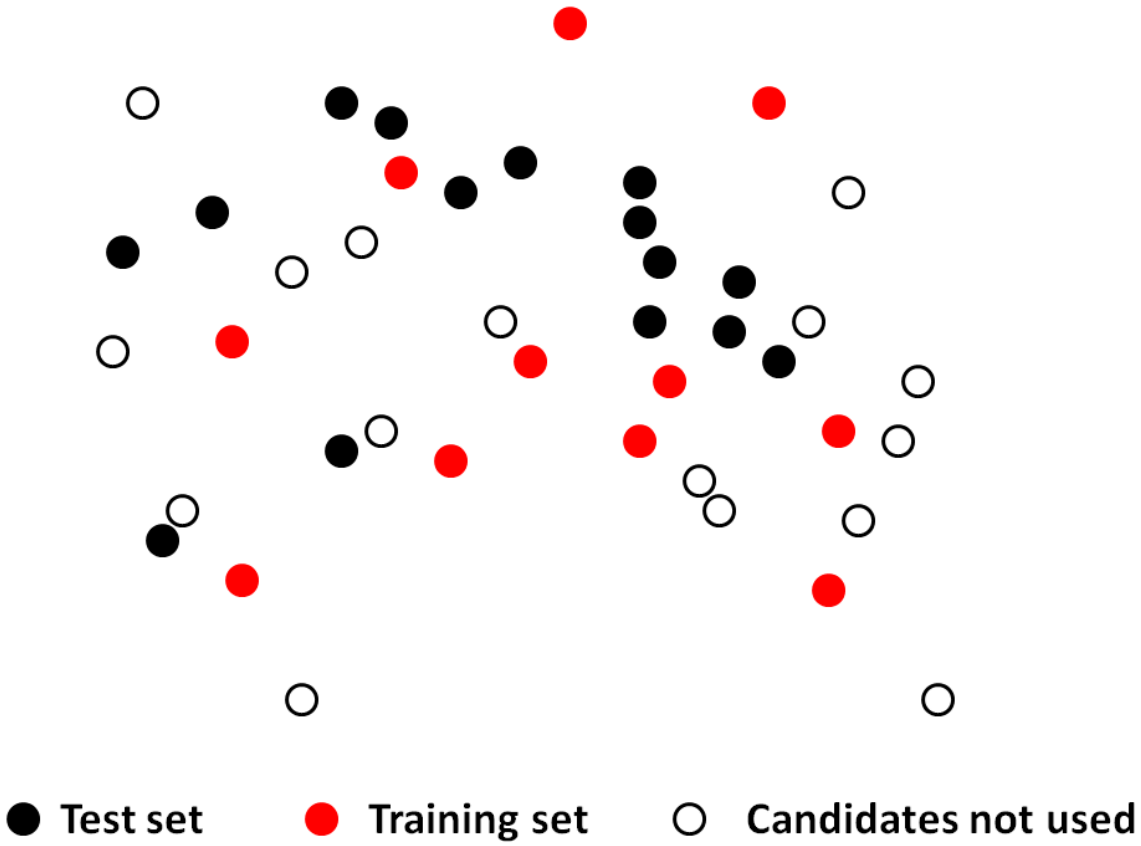
- Tailles des échantillons : 100 TESTS +
  - $n = 500$ Trainings
  - $n = 1000$ Trainings
- $K = 100$, 1000, 5000 ou 10000
- Différentes configurations de corrélation entre les régresseurs
- Nombre de composantes non nulles de $\theta^\star$: 2 ou 100
- $0.50 \leq \frac{\text{Var}(m'_{\text{new}}\theta^\star)}{\text{Var}(Y_{\text{new}})} \leq 0.74$

Comparaison accuracy théorique vs accuracy empirique

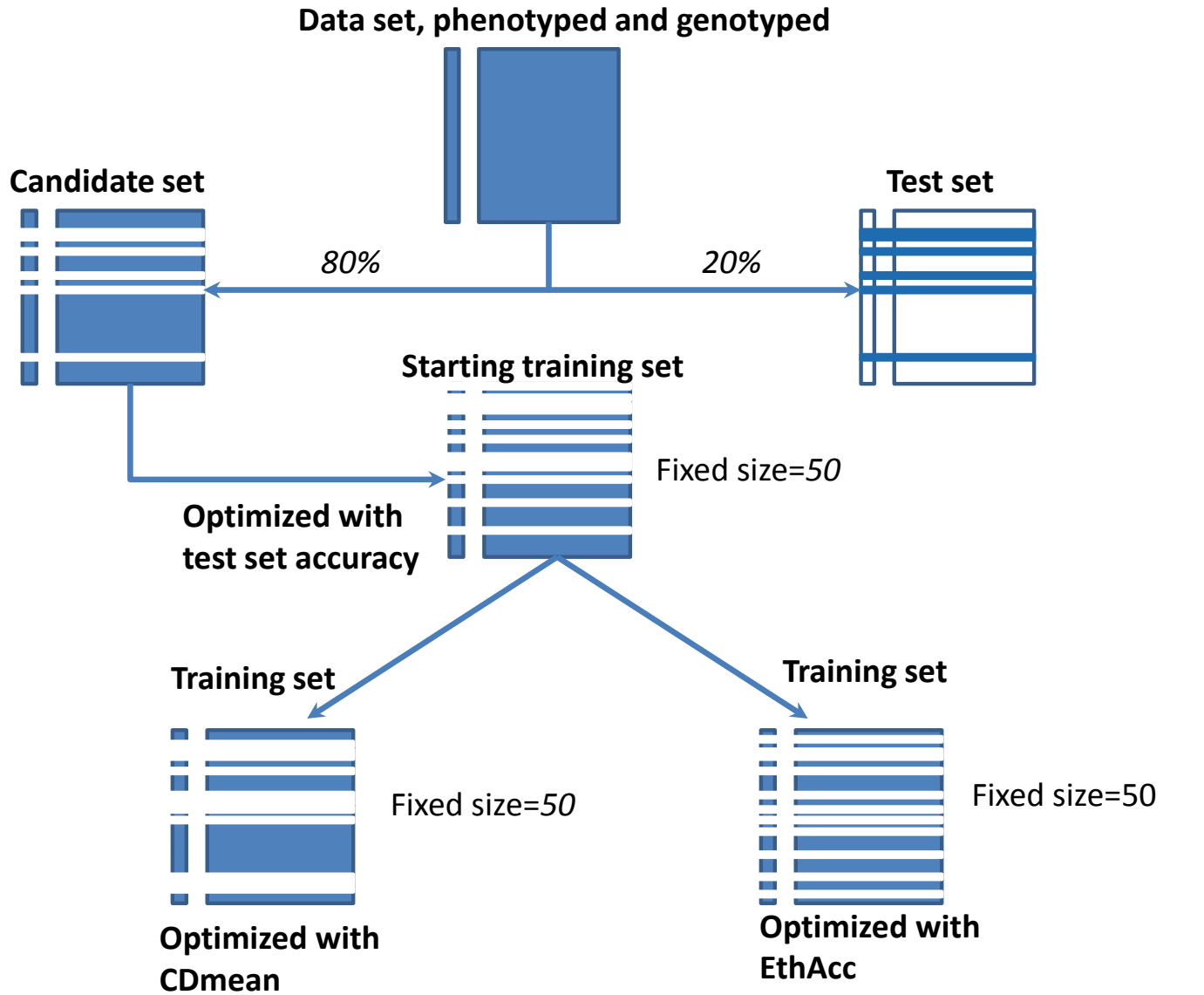


Plan

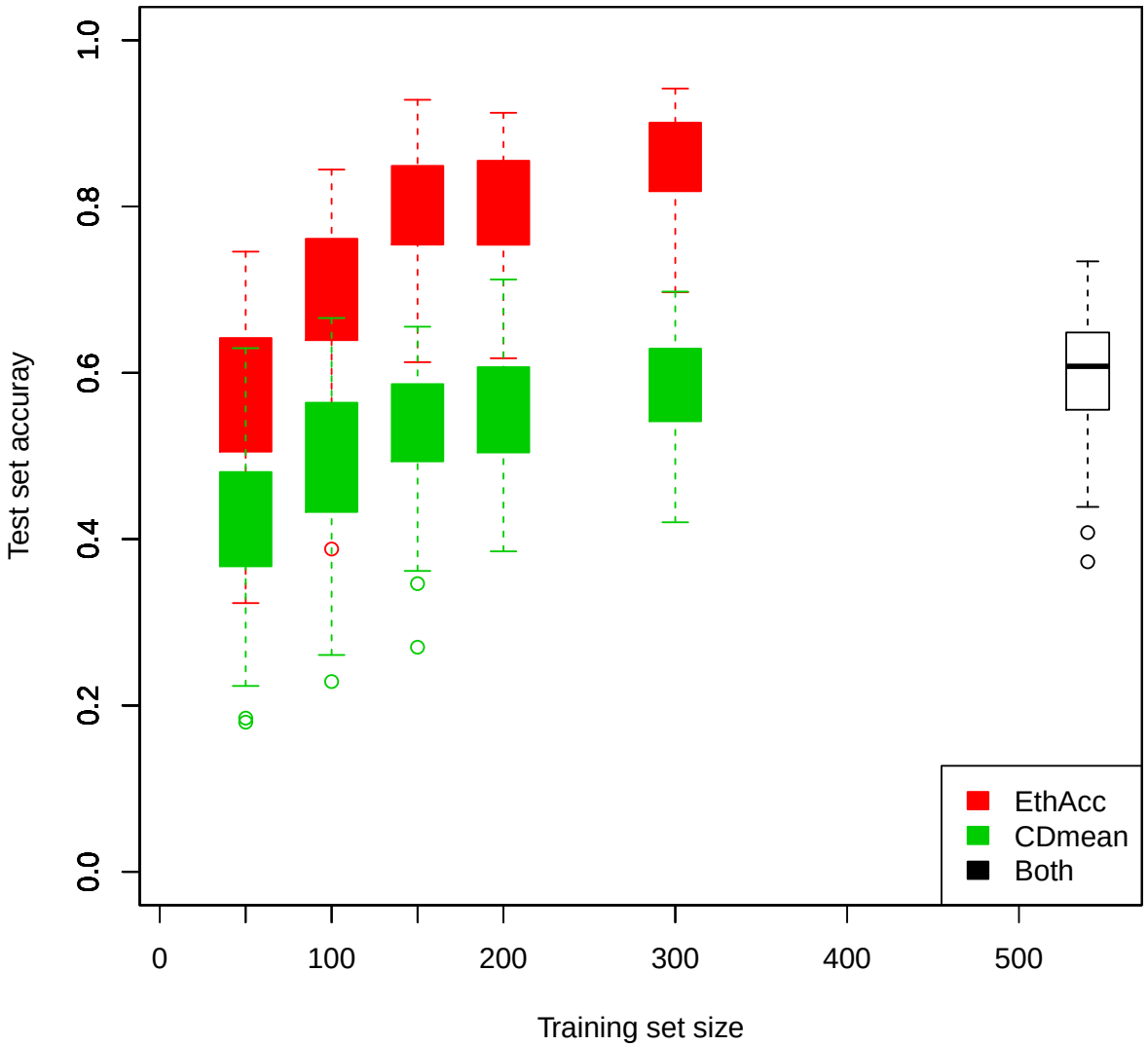
Choix des individus d'apprentissage afin d'effectuer de bonnes prédictions



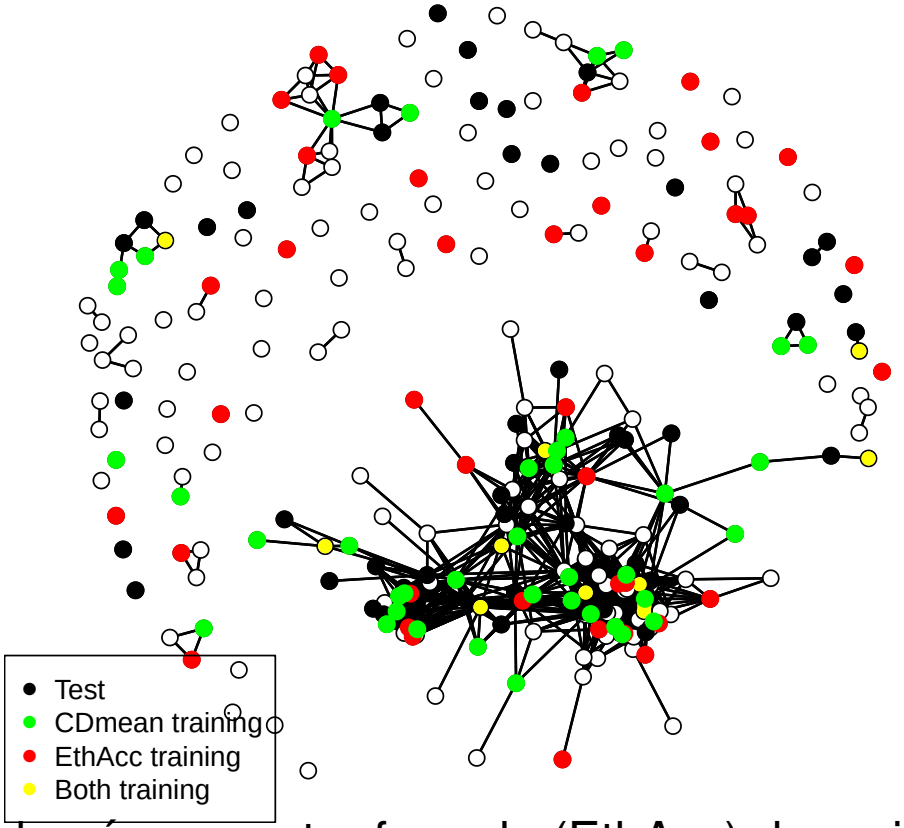
Choix des individus d'apprentissage : Notre approche (EthAcc) versus Modéle mixte (CDmean)



Choix des individus d'apprentissage chez la betterave ($K = 2000$, $n = 500$)



Intérêt de notre approche sur un cas extrême



- Optimisation basée sur notre formule (EthAcc) donnait une accuracy de 0.76
- Optimisation basée sur le modèle mixte (CDmean) donnait une accuracy de 0.07

Plan

Retour sur notre étude de l'accuracy

Critère d'accuracy (i.e. précision de la prédiction)

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{\text{new}}\right)\text{Var}\left(Y_{\text{new}}\right)}}$$

On introduit les notations suivantes :

$$A_1 := \beta' \ \text{Var}\left(m_{\text{new}}\right) M' V^{-1} M\theta^\star \quad , \quad A_2 := \sigma_e^2 \mathbb{E}\left(\left\| m'_{\text{new}} M' V^{-1} \right\|^2\right)$$

$$A_3 := \beta' M' V^{-1} M \text{Var}\left(m_{\text{new}}\right) M' V^{-1} M\beta \quad , \quad A_4 := \text{Var}\left(m'_{\text{new}}\theta^\star\right) + \sigma_e^2.$$

Ainsi,

$$\rho = \frac{A_1}{\left(A_2 + A_3\right)^{1/2}\left(A_4\right)^{1/2}}.$$

Décomposition SVD

Décomposition SVD de M

$$M = U\ D\ W'$$

où

- D matrice diagonale de taille $r \times r$, de plein rang, avec $d_1, \ldots, d_r$ éléments diagonaux
- U matrice de taille $n \times r$, telle que $U'U = I_r$
- W matrice de taille $K \times r$, telle que $W'W = I_r$

A propos de la regression Ridge (Shao et Deng, Annals of Stat 2012)

- WW' est une matrice de projection sur l'espace engendré par les lignes de M
- le projection de $\hat{\theta}$ sur cet espace est encore $\hat{\theta}$

$$\begin{aligned}\hat{\beta} &= WW'\hat{\theta} \\ &= WW'M'\left(MM' + \lambda I_n\right)^{-1} Y \\ &= WW'WDU'\left(MM' + \lambda I_n\right)^{-1} Y \\ &= WDU'\left(MM' + \lambda I_n\right)^{-1} Y \\ &= M'\left(MM' + \lambda I_n\right)^{-1} Y \\ &= \hat{\theta}\end{aligned}$$

- $\beta^\star = WW'\theta^\star$, i.e. la projection de $\theta^\star$ sur cet espace, est ce que nous sommes en mesure d'estimer. $\beta^\star$ pas forcément sparse.
- $\|\beta^\star\| \leq \|\theta^\star\|$. Composantes de $\beta^\star$ très petites.

Résultat sur l'accuracy (i.e. précision de la prédiction)

Théorème (Rabier Mangin Grusea, Scand. J. Stat. 2018)

Soit $\Sigma = Var(m_{\text{new}})$ *la matrice de covariance de taille* $K \times K$*. De plus, on suppose que* M *est connu et que* e, m_{new} *et* e_{new} *sont aléatoires. Alors, on a*

$$\rho = \frac{A_1}{(A_2 + A_3)^{1/2} (A_4)^{1/2}}$$

où

$$A_1 = \sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} \, \theta^{\star\prime} \, \Sigma \, W^{(s)} W^{(s)\prime} \theta^\star \quad , \quad A_4 = \theta^{\star\prime} \Sigma \theta^\star + \sigma_e^2 \, ,$$

$$A_2 = \sigma_e^2 \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \, \mathbb{E}\left(\left\| W^{(s)} W^{(s)\prime} m_{\text{new}} \right\|^2 \right) \, ,$$

$$A_3 = \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} \, W^{(s)} W^{(s)\prime} \theta^\star \right)' \Sigma \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} \, W^{(s)} W^{(s)\prime} \theta^\star \right) .$$

Si on estime l'accuracy ...(TEST et Trainings issus de la même distribution de probabilité)

Théorème (Rabier Mangin Grusea, Scand. J. Stat. 2018)

Supposons que $m_1, \ldots, m_n$ et m_{new} sont indépendantes et identiquement distribuées (i.i.d.). De plus, supposons que $m_1, \ldots, m_n$ ont été observées (i.e. M est connue), et que e, m_{new} et e_{new} sont aléatoires. Alors, une estimation de l'accuracy est la suivante

$$\hat{\rho} = \frac{\widehat{A_1}}{\left(\widehat{A_2} + \widehat{A_3}\right)^{1/2} \left(\widehat{A_4}\right)^{1/2}},$$

où

$$\widehat{A_1} = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^4}{d_s^2 + \lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2 \quad , \quad \widehat{A_2} = \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2 + \lambda)^2}$$

$$\widehat{A_3} = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^6}{(d_s^2 + \lambda)^2} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2 \quad , \quad \widehat{A_4} = \frac{1}{n} \sum_{s=1}^{r} d_s^2 \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2 + \sigma_e^2.$$

Convergence de $\hat{\rho}$ vers ρ^{oracle} lorsque $n \to +\infty$ et $K \to +\infty$

Valeurs singulières

- $d_1 \geq d_2 \geq \ldots \geq d_r > 0$ valeurs singulières de M
- $d_1^2 \sim n^{\psi}$ with $0 < \psi \leq 1$
- $d_r^2 \sim n^{\eta}$ with $\eta \leq \psi \leq 1$ et η et ψ ne dépendant pas de n.

Signal (inspiré de Shao and Deng 2012, et de Fan and Lv 2008)

- $\|WW'\theta^{\star}\|^2 \sim n^{2\tau}$ with $\tau < \eta$ et τ ne dépendant pas de n.

Paramètre de régularisation

- $\lambda \to \infty$ et $\lambda = o\left(d_1^2\right)$

Liens valeurs singulières / paramètre de régularisation

- Ω_1, Ω_2 et Ω_3 désignent les ensembles suivant :

$$\Omega_1 := \left\{ s \,\middle|\, \lambda = o(d_s^2) \right\}, \quad \Omega_2 := \left\{ s \,\middle|\, d_s^2 \sim \frac{1}{C_s}\lambda \text{ with } C_s > 0 \right\},$$

$$\Omega_3 := \left\{ s \,\middle|\, d_s^2 = o(\lambda) \right\}.$$

Convergence de $\hat{\rho}$ vers ρ^{oracle} lorsque $n \to +\infty$ et $K \to +\infty$

Quelques conditions supplémentaires :

- (C1) $\frac{n^{2\tau}}{r} \sum_{s \in \Omega_1} d_s^2 \to +\infty$, (C2) $\sum_{s \in \Omega_3} d_s^2 = o(\lambda)$
- (C3) $\sum_{s \in \Omega_3} d_s^4 = o(\lambda^2)$, (C4) $n^{2\tau}/r = o(1/\lambda)$, i.e. $\lambda = o(r/n^{2\tau})$
- (C5) $\#\Omega_1 = O(1)$, (C6) $\#\Omega_2 = O(1)$

où $\#\Omega$ représente le cardinal de l'ensemble Ω.

Lemma (Convergence vers l'accuracy oracle)

Supposons (C1-C2-C3-C4-C5-C6) et également que le signal est projeté uniformément sur chaque sous espace Vect $\left\{ W^{(s)} \right\}$, *i.e.*

$$\left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2 \sim \frac{n^{2\tau}}{r}, \ s = 1, \ldots, r$$

alors on a $\hat{\rho} \longrightarrow \rho^{oracle} := \sqrt{\frac{Var(m'_{\text{new}} \theta^\star)}{Var(Y_{\text{new}})}}$.

Retour sur notre estimation de l'accuracy ...

- Si échantillons TEST et Trainings issus de la même distribution de probabilité

Théorème

$$\hat{\rho} \geq \frac{\|WW'\theta^\star\|^2 \ \min \frac{d_s^4}{d_s^2+\lambda}}{\sqrt{\sigma_e^2 \ r + \ \|WW'\theta^\star\|^2 \ \max d_s^2} \quad \sqrt{\|WW'\theta^\star\|^2 \ \max d_s^2 + \sigma_e^2}}$$

$$\hat{\rho} = \frac{\sum_{s=1}^{r} \frac{d_s^4}{d_s^2+\lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2}{\left(\sigma_e^2 \ \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} + \sum_{s=1}^{r} \frac{d_s^6}{(d_s^2+\lambda)^2} \|W^{(s)} W^{(s)\prime} \theta^\star\|^2 \right)^{1/2} \left(\sum_{s=1}^{r} d_s^2 \|W^{(s)} W^{(s)\prime} \theta^\star\|^2 + \sigma_e^2 \right)^{1/2}}$$

On aimerait avoir $\sigma_e^2 \ \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2}$ petit, et $\sum_{s=1}^{r} \frac{d_s^4}{d_s^2+\lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2$ grand

Idée : considérer un espace de dimension plus faible

Vers une amélioration de la Ridge

Rappel : $U = \left(U^{(1)}, \ldots, U^{(r)}\right)$ base orthonormale de l'espace engendré par les colonnes de M.

On choisit $\tilde{r}$ colonnes de U. On note $\sigma : \{1, \ldots, \tilde{r}\} \to \{1, \ldots, r\}$

Soit l'estimateur

$$\widetilde{\theta} = M' V^{-1} \widetilde{U}\widetilde{U}' Y \quad \text{où} \quad \widetilde{U} = \left(U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right)$$

où $\widetilde{U}\widetilde{U}' Y$ est la projection de Y sur $Vect\left\{U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right\}$.
On notera $\widetilde{W} = \left(W^{\sigma(1)}, \ldots, W^{\sigma(\tilde{r})}\right)$

$\Rightarrow$ Prédiction et accuracy à l'aide du nouvel estimateur $\widetilde{\theta}$

$$\widetilde{Y}_{\text{new}} = m'_{\text{new}}\widetilde{\theta} \quad , \qquad \widetilde{\rho} = Cor\left(\widetilde{Y}_{\text{new}}, \ Y_{\text{new}}\right) = \frac{\text{Cov}(\widetilde{Y}_{\text{new}}, \ Y_{\text{new}})}{\sqrt{\text{Var}(\widetilde{Y}_{\text{new}})\text{Var}(Y_{\text{new}})}}$$

Accuracy basée sur ce nouvel estimateur

Théorème (Rabier Mangin Grusea, Scand. J. Stat. 2018)

Supposons que $m_1, \ldots, m_n$ et m_{new} sont indépendantes et identiquement distribuées (i.i.d.). De plus, supposons que $m_1, \ldots, m_n$ ont été observées (i.e. M est connue), et que ε, m_{new} et e_{new} sont aléatoires. Alors, une estimation de l'accuracy est la suivante

$$\widehat{\widetilde{\rho}} = \frac{\widehat{\widetilde{A_1}}}{\left(\widehat{\widetilde{A_2}} + \widehat{\widetilde{A_3}}\right)^{1/2} \left(\widehat{\widetilde{A_4}}\right)^{1/2}},$$

où

$$\widehat{\widetilde{A_1}} = \frac{1}{n}\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{d_{\sigma(s)}^2 + \lambda} \left\| W^{(\sigma(s))} W^{(\sigma(s))\prime} \theta^\star \right\|^2 \quad , \quad \widehat{\widetilde{A_2}} = \frac{\sigma_e^2}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2 + \lambda)^2}$$

$$\widehat{\widetilde{A_3}} = \frac{1}{n}\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^6}{(d_{\sigma(s)}^2 + \lambda)^2} \left\| W^{(\sigma(s))} W^{(\sigma(s))\prime} \theta^\star \right\|^2 \quad , \quad \widehat{\widetilde{A_4}} = \widehat{A_4}.$$

Accuracy basée sur ce nouvel estimateur

Lemma

$$\hat{\rho} \geq \frac{\|WW'\theta^\star\|^2 \ \min \frac{d_s^4}{d_s^2+\lambda}}{\sqrt{\sigma_e^2 \ r + \ \|WW'\theta^\star\|^2 \ \max d_s^2} \quad \sqrt{\|WW'\theta^\star\|^2 \ \max d_s^2 + \sigma_e^2}}$$

$$\widehat{\widetilde{\rho}} \geq \frac{\left\|\widetilde{W}\widetilde{W}'\theta^\star\right\|^2 \ \min_{1\leq s\leq \tilde{r}} \frac{d_{\sigma(s)}^4}{d_{\sigma(s)}^2+\lambda}}{\sqrt{\sigma_e^2 \ \tilde{r} + \ \left\|\widetilde{W}\widetilde{W}'\theta^\star\right\|^2 \ \max_{1\leq s\leq \tilde{r}} d_{\sigma(s)}^2} \quad \sqrt{\|WW'\theta^\star\|^2 \ \max d_{\sigma(s)}^2 + \sigma_e^2}}$$

Dans quelles conditions améliore-t-on l'accuracy ?

- Estimateur Ridge $\hat{\theta}$ basé sur toutes les colonnes de U
  - accuracy $\hat{\rho}$, prédiction $\hat{Y}_{\text{new}}$
- Nouvel estimateur $\widetilde{\theta}$ basé sur $\tilde{r}$ colonnes de U
  - accuracy $\tilde{\rho}$, prédiction $\widetilde{Y}_{\text{new}}$
- Complémentaire $\vec{\theta}$ de notre nouvel estimateur basé sur les $r - \tilde{r}$ colonnes restantes de U
  - accuracy $\vec{\rho}$, prédiction $\vec{Y}_{\text{new}}$

Notations :

$$\widehat{A_1} = \widehat{\text{Cov}}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right)\ ,\ \widehat{A_2} + \widehat{A_3} = \widehat{\text{Var}}\left(\hat{Y}_{\text{new}}\right)\ ,\ \widehat{A_4} = \widehat{\text{Var}}\left(Y_{\text{new}}\right)$$

$$\widehat{\widetilde{A_1}} = \widehat{\text{Cov}}\left(\widetilde{Y}_{\text{new}},\ Y_{\text{new}}\right)\ ,\ \widehat{\widetilde{A_2}} + \widehat{\widetilde{A_3}} = \widehat{\text{Var}}\left(\widetilde{Y}_{\text{new}}\right)\ ,\ \widehat{\widetilde{A_4}} = \widehat{A_4} = \widehat{\text{Var}}\left(Y_{\text{new}}\right)$$

...

Les 3 configurations possibles (résultat non asymptotique)

1. On a $\hat{\tilde{\rho}} \geq \hat{\rho}$ si et seulement si

$$\frac{\widehat{\mathrm{Cov}}\left(\tilde{Y}_{\mathrm{new}}, Y_{\mathrm{new}}\right)}{\widehat{\mathrm{Cov}}\left(\vec{Y}_{\mathrm{new}}, Y_{\mathrm{new}}\right)} \geq \frac{\widehat{\mathrm{Var}}\left(\tilde{Y}_{\mathrm{new}}\right)}{\widehat{\mathrm{Var}}\left(\vec{Y}_{\mathrm{new}}\right)} \left(1 + \sqrt{1 + \frac{\widehat{\mathrm{Var}}\left(\vec{Y}_{\mathrm{new}}\right)}{\widehat{\mathrm{Var}}\left(\tilde{Y}_{\mathrm{new}}\right)}}\right).$$

   Dans ce cas, nous avons aussi $\hat{\tilde{\rho}} \geq \hat{\vec{\rho}}$.

2. On a $\hat{\vec{\rho}} \geq \hat{\rho}$ si et seulement si

$$\frac{\widehat{\mathrm{Cov}}\left(\tilde{Y}_{\mathrm{new}}, Y_{\mathrm{new}}\right)}{\widehat{\mathrm{Cov}}\left(\vec{Y}_{\mathrm{new}}, Y_{\mathrm{new}}\right)} \leq \sqrt{1 + \frac{\widehat{\mathrm{Var}}\left(\tilde{Y}_{\mathrm{new}}\right)}{\widehat{\mathrm{Var}}\left(\vec{Y}_{\mathrm{new}}\right)}} - 1.$$

   Dans ce cas, nous avons aussi $\hat{\vec{\rho}} \geq \hat{\tilde{\rho}}$.

3. On a $\hat{\rho} \geq \hat{\tilde{\rho}}$ and $\hat{\rho} \geq \hat{\vec{\rho}}$ si et seulement si

$$\sqrt{1 + \frac{\widehat{\mathrm{Var}}\left(\tilde{Y}_{\mathrm{new}}\right)}{\widehat{\mathrm{Var}}\left(\vec{Y}_{\mathrm{new}}\right)}} - 1 \leq \frac{\widehat{\mathrm{Cov}}\left(\tilde{Y}_{\mathrm{new}}, Y_{\mathrm{new}}\right)}{\widehat{\mathrm{Cov}}\left(\vec{Y}_{\mathrm{new}}, Y_{\mathrm{new}}\right)} \leq \frac{\widehat{\mathrm{Var}}\left(\tilde{Y}_{\mathrm{new}}\right)}{\widehat{\mathrm{Var}}\left(\vec{Y}_{\mathrm{new}}\right)} \left(1 + \sqrt{1 + \frac{\widehat{\mathrm{Var}}\left(\vec{Y}_{\mathrm{new}}\right)}{\widehat{\mathrm{Var}}\left(\tilde{Y}_{\mathrm{new}}\right)}}\right).$$

Un exemple où $\theta^\star$ appartient à l'espace engendré par les lignes de M

- $K = 1000$
- $\theta^\star = 0.3W^{(1)} + 0.3W^{(2)} + 0.3W^{(3)}$
- $\tilde{r}$ et les colonnes de U choisies par validation croisée
- 100 TESTS

σ_e^2	n	Méthode	200 générations	400 générations
1	500	$\hat{\rho}$	0.7550	0.6721
		$\hat{\tilde{\rho}}$	0.7810	0.7041
	800	$\hat{\rho}$	0.7487	0.7037
		$\hat{\tilde{\rho}}$	0.7728	0.7312
9	500	$\hat{\rho}$	0.3370	0.2623
		$\hat{\tilde{\rho}}$	0.3809	0.3079
	800	$\hat{\rho}$	0.3317	0.2904
		$\hat{\tilde{\rho}}$	0.3734	0.3330

Deuxième partie :

Quelques autres de mes thématiques de recherches

Ma recherche

1. Statistique en grande dimension
   - Prédiction
   - Régression Ridge (Pénalité L2)
   - Application aux données réelles
2. Statistique des processus
   - Processus Gaussiens et de Chi-Deux
   - Maximum d'un processus / Max Test
   - Modèles de mélange
   - Plans d'expériences
3. Arbres aléatoires
   - Combinatoire
   - Statistique bayésienne (MCMC)
   - Reconciliation entre arbre d'espèces et arbres de gènes
   - Réseaux phylogénétiques

Echantillonnage des extrêmes
Réduction des coûts liés à une expérience

avec Jean-Marc Azaïs

Université Paul Sabatier Toulouse
Institut de Mathématiques de Toulouse

Modèle en l'absence de censure (oracle)

- X : variable aléatoire discrète correspondant au groupe d'appartenance

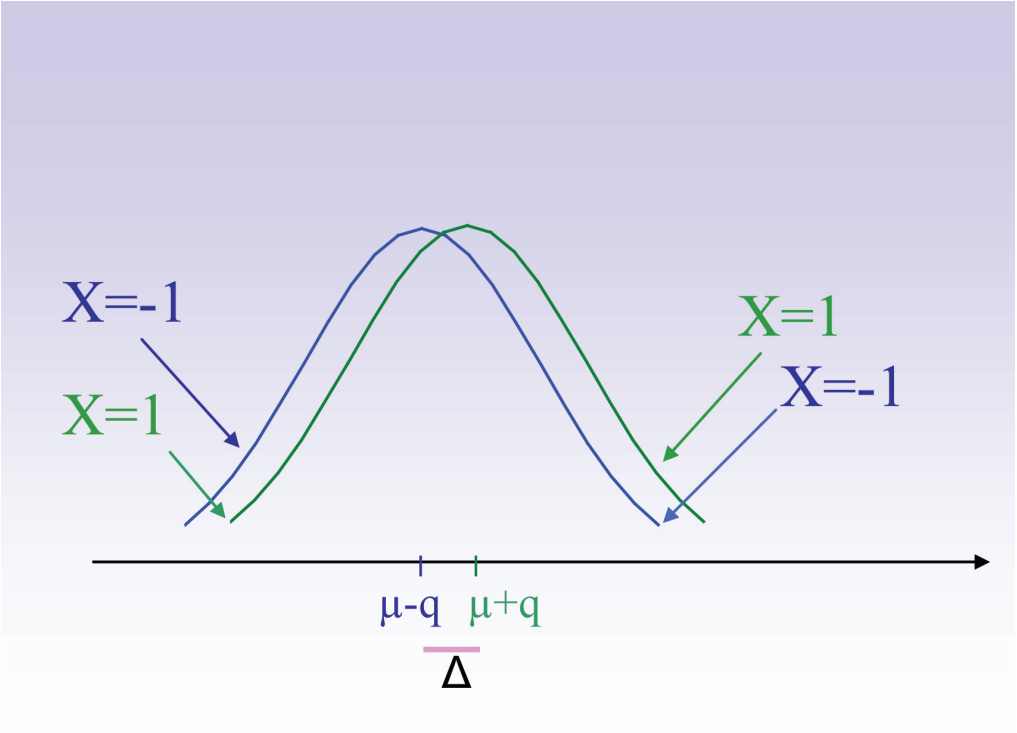
$$X = \begin{cases} -1 & \text{avec probabilité } 1-p \\ 1 & \text{avec probabilité } p \end{cases}$$

On suppose $p \neq \{0, 1\}$

- Y : variable aléatoire continue correspondant à une mesure d'intérêt

$$Y = \mu + q\,X + \sigma\,\varepsilon \qquad \text{où} \quad \varepsilon \sim N(0,1)$$

Modèle en l'absence de censure (oracle)



Distribution de la variable aléatoire Y

Test statistique oracle (μ, q, σ)

- A l'aide de n observations (X_j, Y_j) i.i.d., on souhaite tester :

$$H_0 : \ q = 0 \ \ vs \ \ H_1 : \ q \neq 0$$

On considère une alternative locale $H_a : q = \frac{a}{\sqrt{n}}$

- Test statistique oracle :

$$T = \frac{\sum_{j=1}^{n} \ \frac{1}{p}(Y_j - \overline{Y}) \ 1_{X_j=1} \ - \ \frac{1}{1-p}(Y_j - \overline{Y}) \ 1_{X_j=-1}}{\hat{\sigma} \ \sqrt{\frac{n}{p \ (1-p)}}}$$

$$T \overset{H_0}{\rightarrow} N(0, \ 1) \quad \text{et} \quad T \overset{H_a}{\rightarrow} N\left(\frac{2a \ \sqrt{p \ (1-p)}}{\sigma}, \ 1\right)$$

Echantillonnage des extrêmes

Obtenir l'information groupe d'appartenance (X) coûte cher

On mesure X uniquement pour les individus présentant des Y extrêmes

Le nombre d'individus pour lesquels X est mesuré, afin d'obtenir une puissance donnée, est réduit considérablement à condition que le nombre total d'individus ait été augmenté

Lebowitz et al. (Theoretical and Applied Genetics, 1987)

Questions abordées

- Combien d'individus supplémentaires faut-il pour avoir même puissance qu'en situation oracle ?
- Doit-on collecter les X uniquement pour les individus présentant les Y les plus grands, ou au contraire pour ceux présentant les Y les plus petits, ou bien un mélange des deux ?
- Doit-on conserver les Y non extrêmes dans l'analyse statistique ?

Modèle basé sur les extrêmes

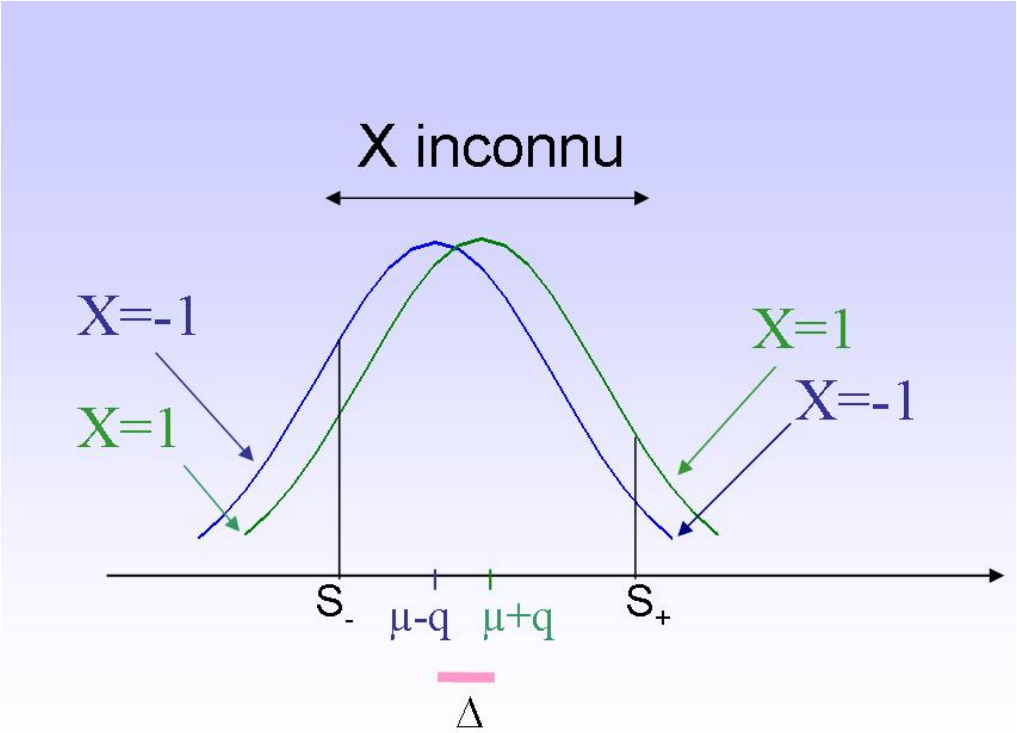
X disponible uniquement lorque Y est extrême

$\Rightarrow$ On n'observe plus X mais $\overline{X}$:

$$\overline{X} = \begin{cases} X & \text{si } \ Y \notin [S_- \ , \ S_+] \\ 0 & \text{sinon} \end{cases}$$

où S_- et S_+ sont deux réels tels que $S_- \leq S_+$.

Modèle basé sur les extrêmes

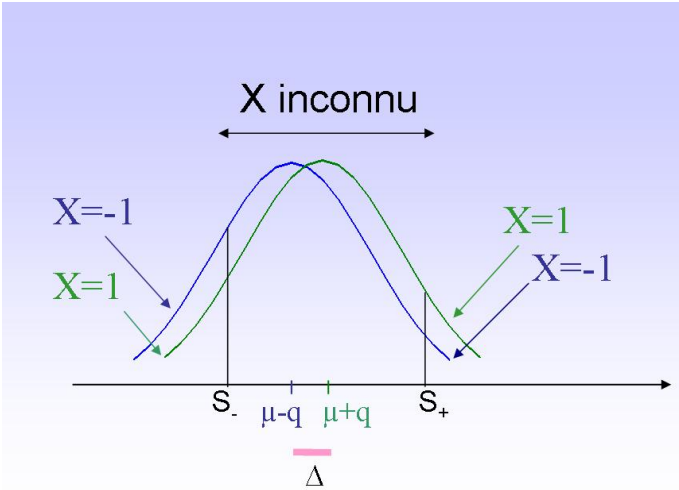


Distribution de la variable aléatoire Y

Comparaison des 3 stratégies

3 stratégies pour l'analyse de données extrêmes :

1. Test de Wald basé sur l'ensemble des Y
2. Comparaison de moyenne basée sur les Y extrêmes
3. Test de Wald basé sur les Y extrêmes



Notations :

$$\gamma = \mathbb{P}_{H_0}\left(Y \notin [S_-; S_+]\right) \quad , \quad \gamma_+ = \mathbb{P}_{H_0}\left(Y > S_+\right]) \quad , \quad \gamma_- = \mathbb{P}_{H_0}\left(Y < S_-\right])$$

$$\gamma_+ + \gamma_- = \gamma$$

A la fois sous H_0 et sous H_a, γ correspond asymptotiquement au pourcentage d'individus pour lesquels X a été collecté

Comparaison des 3 stratégies (μ, q, σ)

Lemme

$$W_1 := \frac{2\sqrt{n}}{\hat{\sigma}^2}\sqrt{\hat{\mathcal{A}}\, p(1-p)}\;\; \hat{q}_1$$

$$T_2 := \sqrt{p(1-p)}\left\{\frac{\sum_{j=1}^n \frac{1}{p}(Y_j-\overline{Y})1_{\overline{X}_j=1} - \frac{1}{1-p}(Y_j-\overline{Y})1_{\overline{X}_j=-1}}{\sqrt{n\,\hat{\mathcal{A}}}}\right\}$$

$$W_3 := \frac{2\sqrt{n}}{\hat{\sigma}^2}\sqrt{\hat{\mathcal{A}}\, p(1-p)}\;\; \hat{q}_3$$

présentent les mêmes lois asymptotiques sous H_0 et sous H_a, à savoir :

$$N(0,\ 1) \quad \textit{et} \quad N\left(\frac{2a\,\sqrt{\mathcal{A}\, p(1-p)}}{\sigma^2},\ 1\right)$$

où $\hat{q}_1$ et $\hat{q}_3$ sont les EMV de q pour les stratégies une et trois, et où

$$\hat{\mathcal{A}} = \frac{1}{n}\sum_{j=1}^n (Y_j-\overline{Y})^2 1_{\overline{X}_j\neq 0} \quad , \quad \overline{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$$

$$\mathcal{A} = \sigma^2\left\{\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\} \quad , \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^n (Y_j-\overline{Y})^2$$

Modèle “Locally Asymptotically Normal” + 3ème lemme de Le Cam

Efficacité au sens de Pitman

A propos des tailles échantillonnales

- n : nombre d'observations dans une expérience oracle
- $n^\star$: nombre d'observations nécessaires dans une experience basée sur les extrêmes afin d'obtenir la même puissance que dans l'expérience oracle

Efficacité au sens de Pitman

$$\kappa = \frac{n}{n^\star}$$

Après calcul, on a

$$\kappa = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})$$

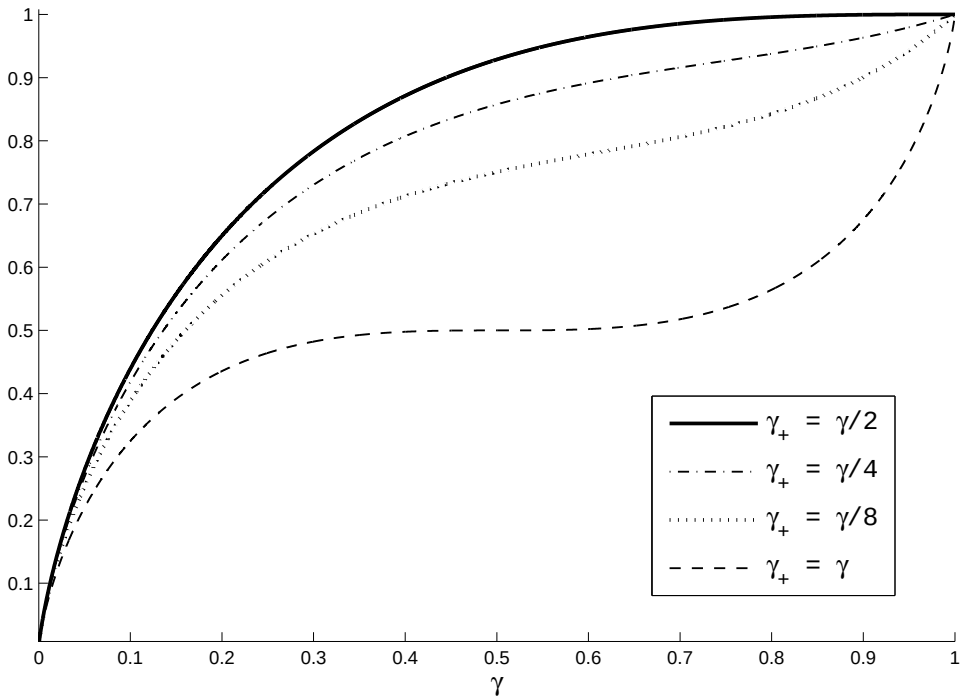
L ' échantillonnage des extrêmes s'avère intéressant uniquement si

$$\gamma n^\star < n$$

En d'autres termes, on doit avoir

$$\gamma < \kappa$$

Efficacité au sens de Pitman, en fonction du pourcentage γ d'observations extrêmes, et en fonction du ratio γ_+/γ



Conclusions sur l'échantillonnage des extrêmes

- $\forall p$, on doit collecter le même pourcentage de données aux deux extrêmes ($\gamma_+/\gamma = 1/2$)
- Les Y non extrêmes n'apportent pas d'information
- Le γ optimal dépend du ratio c_X/c_Y (environ 30% si $c_X/c_Y = 2$)
- Le test de comparaison de moyenne est optimal

	$n = 50$		$n = 100$	
Nombre de tests	W_1	T_2	W_1	T_2
1	0.0020	0.0005	0.0041	0.0005
1000	2.7871	0.1267	5.1131	0.1384

Comparaison en temps de calcul
($q = 0.3$, $p = 1/2$, $\gamma = 0.3$, $\gamma_+/\gamma = 1/2$)

Rabier, JSPI 2014

Statistique bayésienne / Reconstruction de réseaux

avec Vincent Berry, Fabio Pardi et Céline Scornavacca

ISEM, Institut des Sciences de l'Evolution de Montpellier
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique

Réseaux phylogénétiques

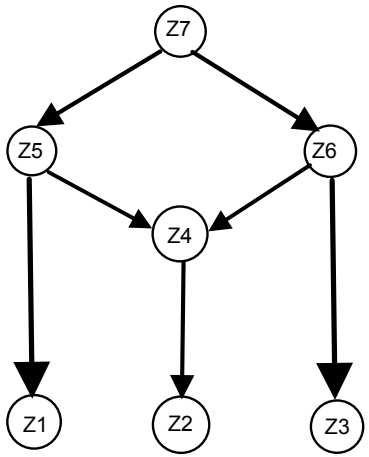
Les réseaux phylogénétiques sont des DAG qui vont nous permettre de détecter des évènements évolutifs :

- hybridations
- introgressions
- transferts horizontaux

Quelques points importants :

- Longueur d'une arête = temps d'évolution
- Dépendance entre noeuds (probabilités conditionnelles ?)
- On cherche à avoir une distribution de réseaux (incertitude sur des clades)
- Plus on collecte de données, plus on est en mesure d'inférer précisément le réseau

Les réseaux phylogénétiques sont des DAG



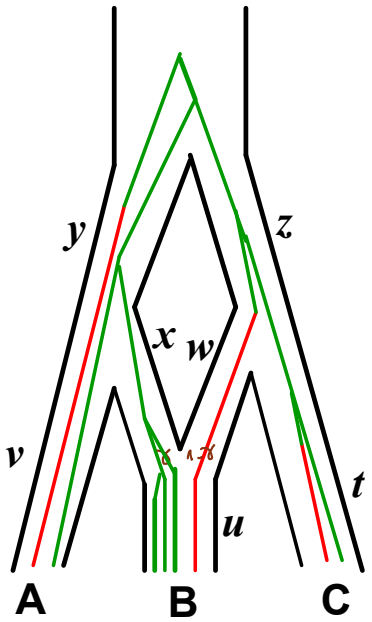
- Noeud 4 : noeud de réticulation
- Noeud 1, 2 et 3 : feuilles du réseau
- Z_k : v.a. correspondant au noeud k
- Loi jointe :

$$\begin{aligned}&\mathbb{P}(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7)\\ &= \mathbb{P}(Z_2 \mid Z_4)\,\mathbb{P}(Z_1, Z_3, Z_4 \mid Z_5, Z_6)\\ &\times \mathbb{P}(Z_5, Z_6 \mid Z_7)\,\mathbb{P}(Z_7)\end{aligned}$$

- Z_1, Z_2, Z_3 ne sont pas de même nature que Z_4, Z_5, Z_6, Z_7
- 2 types d'allèles (rouge/vert)
- N_k : v.a. pour le nombre de lignées au noeud k
- R_k : v.a. pour le nombre d'allèles rouge au noeud k
- $\forall k \in \{4,5,6,7\}$, $Z_k = (N_k, R_k)$
- $\forall k \in \{1,2,3\}$, $Z_k = R_k$
- N_1, N_2 et N_3 sont connus ! ! !
- Data=(Z_1, Z_2, Z_3)

Réseaux phylogénétiques

- Modélisation de l'arbre de locus (backward) :
  - multispecies coalescent
  - modèle de Nakhleh au niveau du noeud de réticulation
- Modélisation des données au SNP (forward)



- mutation (rouge $\leftrightarrow$ vert) : modèle markovien évoluant le long des branches de l'arbre de locus
- u : taux de mutation rouge $\rightarrow$ vert
- v : taux de mutation vert $\rightarrow$ rouge

Une méthode Bayésienne d'inférence de réseaux

- N : réseau phylogénétique (topologie, longueurs de branches, tailles de populations)
- X_i : données pour le SNP i
- G_i : arbre de locus pour le SNP i
- m SNPs

$$\mathbb{P}(N|X_1,\ldots,X_m) \propto \left(\prod_{i=1}^{m}\int_{\psi}\mathbb{P}(X_i|G_i)\mathbb{P}(G_i|S)dG_i\right)P(N)$$
$$\propto \mathbb{P}\left(Data \mid N\right)P(N)$$

SNAPPNet intègre sur tous les arbres de locus (extension de SNAPP, Bryant et al. MBE 2012), à l'aide d'un nouvel algorithme de parcours du réseau

Calcul de la *prior* P(N) par le processus de naissances hybridation

$\Rightarrow$ Markov Chain Monte Carlo (MCMC) afin d'estimer la distribution à posteriori de $\mathbb{P}(N|X_1,\ldots,X_m)$

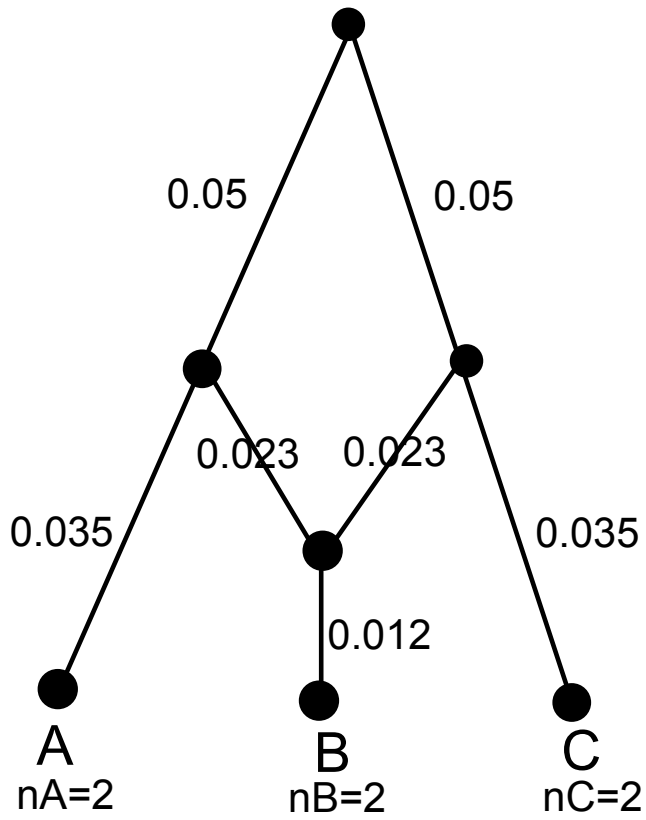
A propos de l'exploration de l'espace des réseaux

Nous cherchons à avoir une distribution de réseaux (incertitude sur des clades)

Cadre du MCMC

- A priori sur le réseau : Processus de Naissance Hybridation
- L'a priori contrôle l'espace des réseaux étudiés
  - Réseaux avec peu de réticulations (faible taux d'hybridation)
  - Réseaux avec beaucoup de réticulations (fort taux d'hybridation)
- Opérateurs de changements topologiques
- Opérateurs de changements de longueurs de branches, etc ...

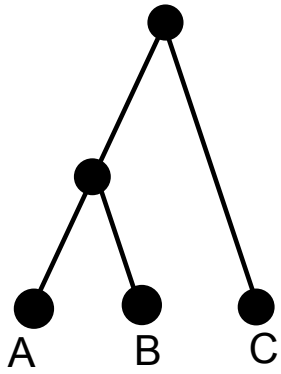
Un exemple sur données simulées



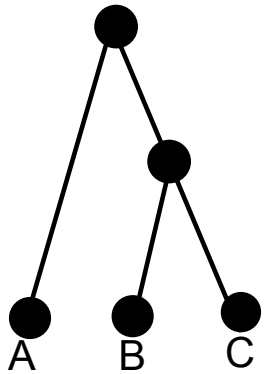
- Longueurs de branches en nombre de mutations par site
- nA=2, nB=2, nC=2
- 1 000 sites ou 10 000 sites
- Tailles de population θ égales à 0.005 ou 0.05
- T : temps de coalescence entre 2 lignées (en mutations par site)
  - si $\theta = 0.005$, alors $\mathbb{E}(T) = 0.005/2 = 0.0025$
  - si $\theta = 0.05$, alors $\mathbb{E}(T) = 0.005/2 = 0.025$

Résultats obtenus par MCMC

- 1 000 sites, $\theta = 0.005$

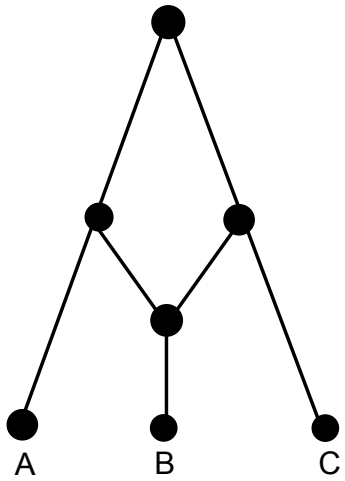


51.94%



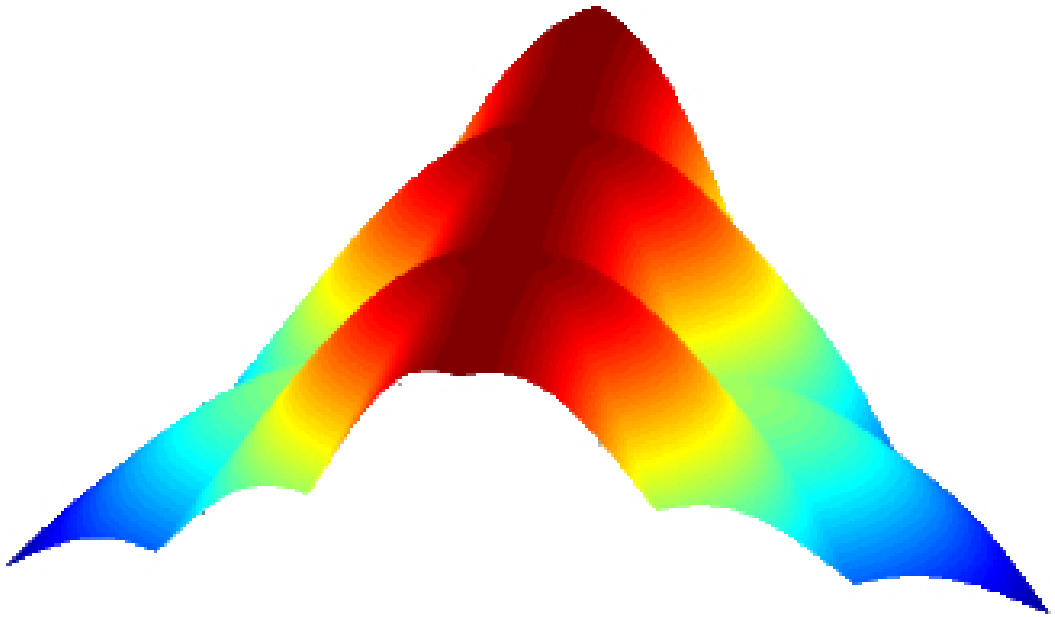
48.06%

- 10 000 sites, $\theta = 0.005$
- 1 000 sites, $\theta = 0.05$
- 10 000 sites, $\theta = 0.05$



100%

Merci de votre attention



Quelques questions sur la régression Ridge

Contexte of Shao and Deng (AOS, 2012)

- Design fixe M
- TEST=Training, i.e. TEST non aléatoire

Théorème 1 de Shao and Deng

- $rank(M) = r_n$
- if $r_n/n \to 0$, erreur de prédiction L2 peut tendre vers 0

$$\frac{1}{n}\mathbb{E}\left\|M\hat{\theta} - M\beta^\star\right\|^2$$

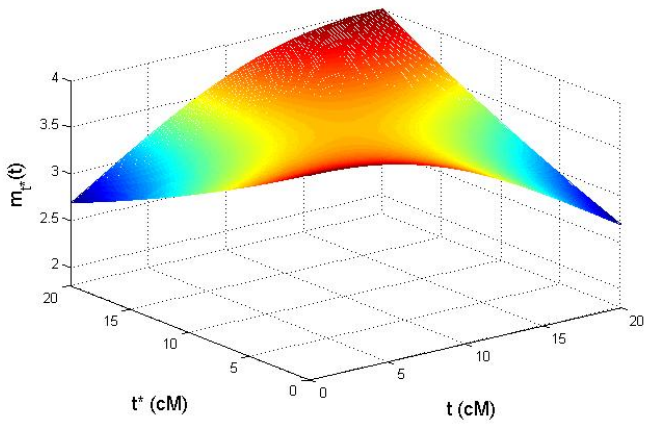
- que se passe-t-il avec des individus TEST aléatoires ?

$$\frac{1}{n}\mathbb{E}\left\|m'_{\text{new}}\hat{\theta} - m'_{\text{new}}\beta^\star\right\|^2$$

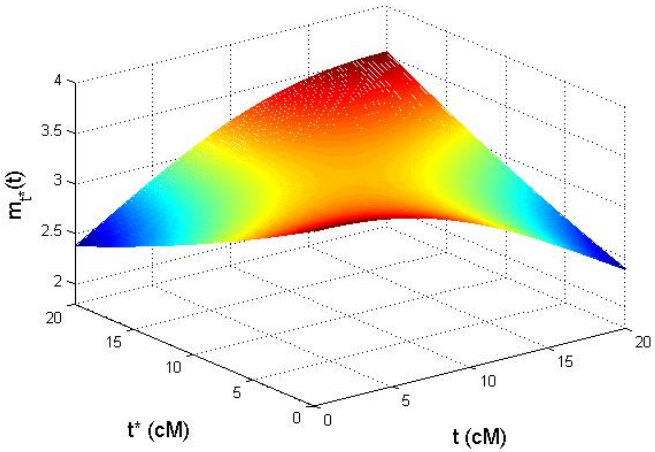
Notre étude :

$$\rho = \frac{\text{Cov}\left(m'_{\text{new}}\hat{\theta},\ Y_{\text{new}}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{\text{new}}\right)\text{Var}\left(Y_{\text{new}}\right)}}$$

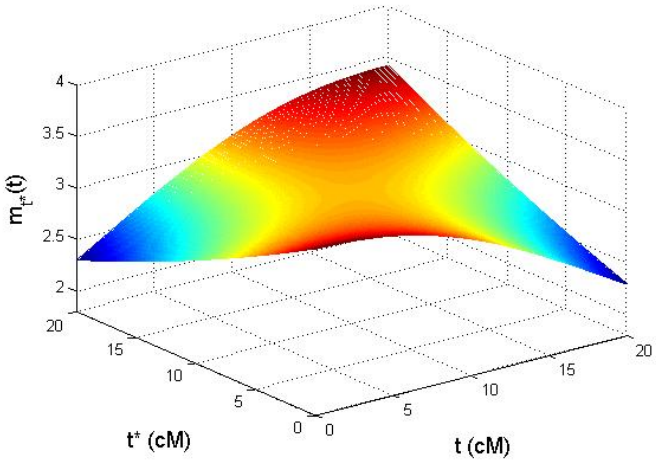
Fonction moyenne d'un processus Gaussien étudié en génomique



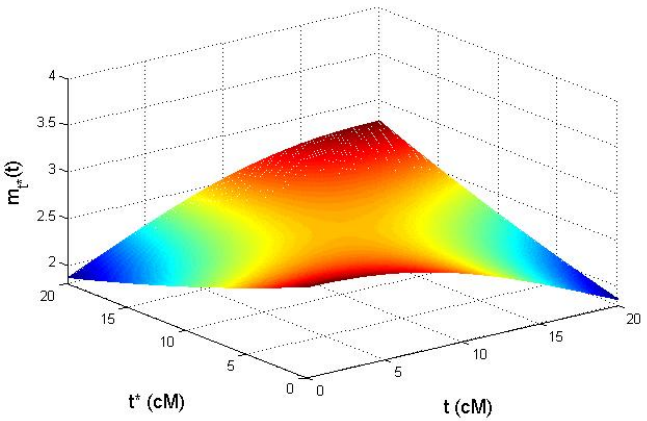
$\gamma = 1$



$\gamma = 0.3$, $\gamma_+ = \gamma/2$

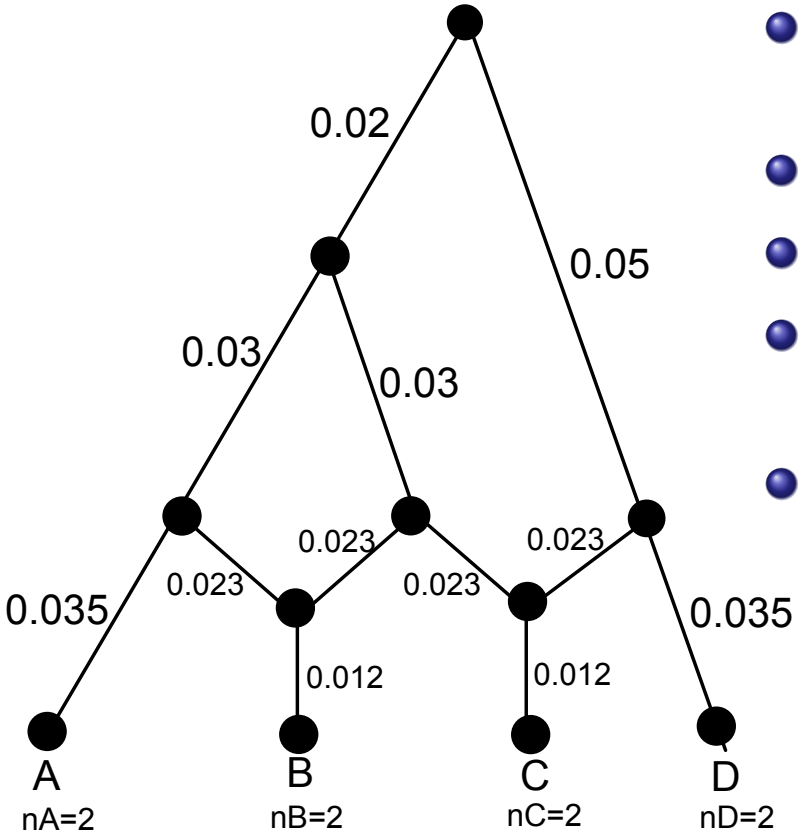


$\gamma = 0.3$, $\gamma_+ = 3\gamma/4$



$\gamma = 0.3$, $\gamma_+ = \gamma$

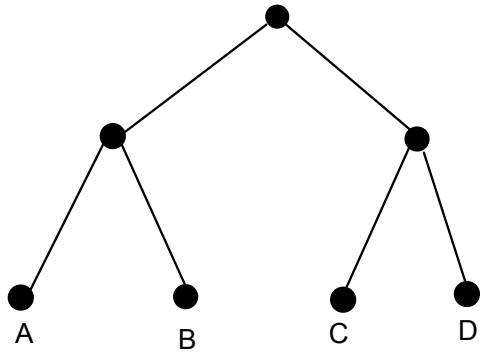
Un réseau avec 2 réticulations



- Longueurs de branches en nombre de mutations par site
- nA=2, nB=2, nC=2, nD=2
- 1 000 sites ou 10 000 sites
- Tailles de population θ égales à 0.005 ou 0.05
- T : temps de coalescence entre 2 lignées (en mutations par site)
  - si $\theta = 0.005$, alors $\mathbb{E}(T) = 0.005/2 = 0.0025$
  - si $\theta = 0.05$, alors $\mathbb{E}(T) = 0.005/2 = 0.025$

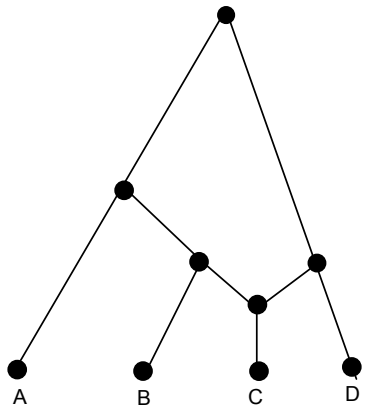
Résultats obtenus par MCMC

- 1 000 sites

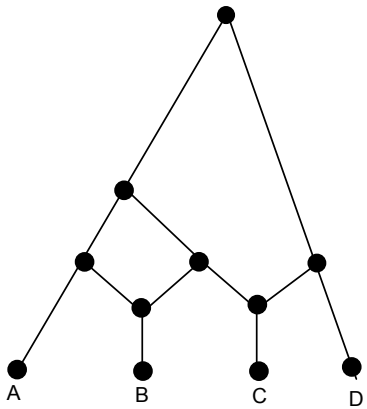


98.44% pour theta=0.05
99.90% pour theta=0.005

- 10 000 sites



100% pour theta=0.005



98.99% pour theta=0.05

Fonction de coûts

La taille echantillonnale requise pour atteindre une puissance β, en considérant un test de niveau α, est la quantité $n_{\alpha,\beta}$ vérifiant :

$$n_{\alpha,\beta} = \frac{\sigma^4 \, (z_\alpha - z_\beta)^2}{4 \, q^2 \, \mathcal{A} \, p(1-p)}$$

Soit :

- c_X (resp. c_Y) : coût pour collecter X (resp. Y)
- C ratio c_X / c_Y

Afin de trouver le γ optimal, nous devons minimiser la fonction suivante

$$F(\gamma) = n_{\alpha,\beta} \; \gamma \; c_X \; + \; n_{\alpha,\beta} \; c_Y = \frac{\sigma^4 \, (z_\alpha - z_\beta)^2 \; c_Y \; (\gamma \; C \; + \; 1)}{4 \, q^2 \, \mathcal{A} \, p(1-p)}$$

A titre d'exemple, si $\gamma_+ / \gamma = 1/2$:

- Si $C = 5$, $\gamma_{\text{opt}} \approx 0.2$
- Si $C = 2$, $\gamma_{\text{opt}} \approx 0.3$

γ optimal en fonction du ratio $C = c_X/c_Y$