Prédiction génomique en grande dimension : densité de marqueurs nécessaire pour une prédiction fiable en sélection génomique

Charles-Elie Rabier, Simona Grusea

Institut Montpelliérain Alexander Grothendieck
Key Initiative MUSE Data & Life Sciences
Institut de Mathématiques de Toulouse
Institut National des Sciences Appliquées de Toulouse

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

Densité de marqueurs requise pour une bonne prédiction

- maïs (Zhang et al, Heredity 2015)
  - 58000 marqueurs nécessaires pour un trait complexe
  - 200 marqueurs nécessaires pour un trait simple
- ray-grass (notre étude, Plos One 2016) : 24957 marqueurs densité insuffisante pour couvrir le génome entier (2.7 Gb)
- café (Ferrao et al, Heredity 2018) : prédictions basées sur 4000 marqueurs $\sim$ 35000 marqueurs
- aquaculture (Kriaridou et al, Frontiers in Genetics 2020) : prédictions basées sur 1000 marqueurs $\sim$ grande densité de marqueurs

Modèles statistiques

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
ne contient pas forcément les vrais régresseurs ...

Autrement dit, chaque colonne de $M^\star$ n'est pas forcément une colonne de M

Echantillon de validation + critère d'accuracy

- Soit un individu TEST noté new

$$Y_{\text{new}} = m^{\star\prime}_{\text{new}} \; \theta^\star + e_{\text{new}} \quad \text{où} \quad e_{\text{new}} \sim N(0, \sigma^2_e)$$

et $m^\star_{\text{new}}$ vecteur de mesures de l'individu new

- Prédiction de la variable continue Y_{new}

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

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

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

avec m_{new} et $m^\star_{\text{new}}$ aléatoires, M fixe

Composante essentielle dans l'équation du sélectionneur (cf. Lynch and Walsh, 1998)

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

Situation oracle

Situation oracle :

$$\theta^\star \text{ connu et donc } \hat{Y}_{\text{new}} = m_{\text{new}}^{\star\prime}\theta^\star$$

Accuracy oracle :

$$\rho^{oracle} := \text{Cor}\left(m_{\text{new}}^{\star\prime}\theta^\star,\ Y_{\text{new}}\right) = \sqrt{\frac{\text{Var}(m_{\text{new}}^{\star\prime}\theta^\star)}{\text{Var}(Y_{\text{new}})}} = h$$

Dans le meilleur des cas, la précision de la prédiction est
la racine carré de l'héritabilité du trait

A propos de l'accuracy dans le cadre de la Ridge

Le prédicteur s'écrit $\hat{Y}_{\text{new}} = m'_{\text{new}} \left(M'M + \lambda I_K\right)^{-1} M'Y$

On introduit les notations suivantes :

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

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

Pour la Ridge, on a

$$\rho = \mathrm{Cor}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right) = \frac{A_1}{\left(A_2 + A_3\right)^{1/2}\left(A_4\right)^{1/2}}.$$

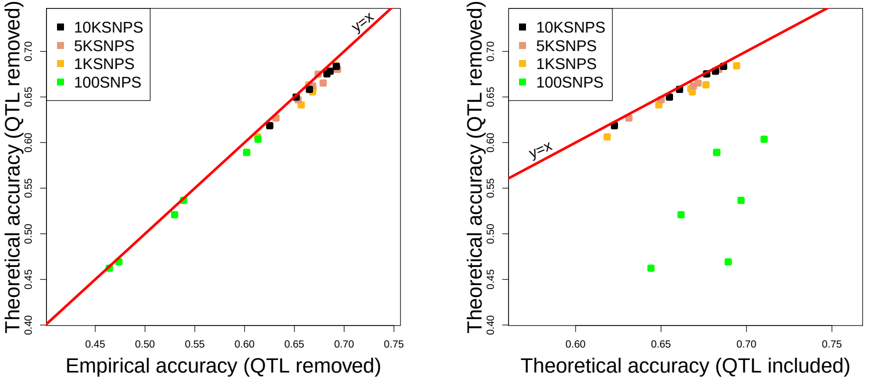
On a aussi

$$\mathbb{E}\left\{\left(\hat{Y}_{\text{new}} - m^{\star\prime}_{new}\theta^{\star}\right)^2\right\} = A_2 + A_3 + A_4 - 2A_1\ .$$

Illustration 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
- $\theta^\star$ vecteur de taille 2 ou 100 (i.e. $Q = 2$ ou $Q = 100$)
- $0.50 \leq \frac{\mathrm{Var}(m_{\mathrm{new}}^{\star\prime}\theta^\star)}{\mathrm{Var}(Y_{\mathrm{new}})} \leq 0.74$

Les QTLs ne sont pas situés sur les marqueurs (DL imparfait)



Décompositions SVD utiles pour notre étude

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 $W'W = I_r$

De la même manière,

$$M^\star = U^\star\ D^\star\ W^{\star\prime}$$

où

- $D^\star$ matrice diagonale de taille $r^\star \times r^\star$, de plein rang, avec $d_1^\star, \ldots, d_{r^\star}^\star$ éléments diagonaux
- $U^\star$ matrice de taille $n \times r^\star$, telle que $U^{\star\prime}U^\star = I_{r^\star}$
- $W^\star$ matrice de taille $Q \times r^\star$, telle que $W^{\star\prime}W^\star = I_{r^\star}$

A propos de la régression Ridge

- 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 $\hat{\theta}$

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

Une bonne lecture : Shao et Deng (Annals of Stat 2012)

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

Théorème (R. et Grusea, JRSS C 2021)

Supposons que $m_1, \ldots, m_n$ et m_{new} sont indépendantes et identiquement distribuées (i.i.d.). De la même facon, supposons que $m_1^\star, \ldots, m_n^\star$ et $m_{new}^\star$ sont i.i.d. De plus, supposons que $m_1, \ldots, m_n$, $m_1^\star, \ldots, m_n^\star$, ont été observées (i.e. M et $M^\star$ sont connues), 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

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

$$\widehat{A_3} = \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}\left\|U^{(s)}U^{(s)\prime}M^\star\theta^\star\right\|^2 \quad , \quad \widehat{A_4} = \frac{1}{n}\sum_{s=1}^{r^\star}d_s^{\star 2}\left\|W^{\star(s)}W^{\star(s)\prime}\theta^\star\right\|^2 + \sigma_e^2.$$

On peut retrouver les resultats avec les colonnes de $M^\star$ présentes dans M

Si les colonnes de $M^\star$ sont présentes dans M ...

Il suffit de remplacer $M^\star\theta^\star$ par $M\widetilde{\theta}^\star$ avec $\widetilde{\theta}^\star$ vecteur sparse de taille K contenant les composantes de $\theta^\star$

Théorème (R. Mangin Grusea, Scand. J. Stat. 2019)

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

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

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

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 dèpendant pas de n.

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

- $\left\| WW'\widetilde{\theta}^{\star} \right\|^2 \sim n^{2\tau}$ with $\tau < \eta$ et τ ne dèpendant 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 :

$$\Omega_1 := \left\{ s \middle| \lambda = o(d_s^2) \right\}, \ \Omega_2 := \left\{ s \middle| d_s^2 \sim \frac{1}{C_s}\lambda \text{ avec } C_s > 0 \right\},$$
$$\Omega_3 := \left\{ s \middle| d_s^2 = o(\lambda) \right\}.$$

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.*

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

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

On revient au cas où les colonnes de $M^\star$ pas forcément présentes dans M ...

$\widehat{A_1}$ et $\widehat{A_3}$ se réecrivent de la manière suivante :

$$\widehat{A_1} = \frac{1}{n} \sum_{s=1}^{r} \theta^{\star\prime} \frac{d_s^2}{d_s^2 + \lambda} \sum_{\ell=1}^{r^\star} W^{\star(\ell)} d_\ell^\star U^{\star(\ell)\prime} U^{(s)} \sum_{j=1}^{r^\star} d_j^\star U^{(s)\prime} U^{\star(j)} W^{\star(j)\prime} \theta^\star \ ,$$

$$\widehat{A_3} = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2 + \lambda)^2} \left(\sum_{\ell=1}^{r^\star} d_\ell^\star U^{(s)\prime} U^{\star(\ell)} W^{\star(\ell)\prime} \theta^\star \right)^2 .$$

on discute le lien entre les valeurs singulières $d_s^\star$, d_s, λ
et projection du signal sur sous espaces associés

On fait intervenir une partition $\Omega_1^\star$, $\Omega_2^\star$, $\Omega_3^\star$ de $\{1, \dots, r^\star\}$
On montre que pour n large, $\hat{\rho}$ se comporte comme $\sqrt{\xi(n)}\ \rho^{oracle}$

$1 - \xi(n)$: pourcentage de la norme L^2 de $\boldsymbol{U}^{\star(\ell)}$ que l'on n'arrive pas à capter
...

Quelques définitions

Pour chaque $\ell \in \{1, \ldots, r^\star\}$, on définit les ensembles Ω_k^ℓ, $k = 1, 2, 3$:

$$\Omega_k^\ell := \left\{ s \in \Omega_k \mid \left\| U^{(s)} U^{(s)\prime} U^{\star(\ell)} \right\|^2 \neq 0 \right\}.$$

En d'autres termes, on assume que la projection de $U^{\star(\ell)}$ sur $Vect\left\{U^{(1)}, \ldots, U^{(r)}\right\}$ est éparpillée sur les sous espaces $\underset{s \in \Omega_1^\ell}{Vect}\left\{U^{(s)}\right\}$, $\underset{s \in \Omega_2^\ell}{Vect}\left\{U^{(s)}\right\}$, et $\underset{s \in \Omega_3^\ell}{Vect}\left\{U^{(s)}\right\}$.

Pour $k = 1, 2, 3$, on impose $\Omega_k^\ell \cap \Omega_k^{\ell'} = \emptyset$, $\forall \ell \neq \ell'$.
Autrement dit, un "s" donné ne peut pas cibler différents "ℓ".

Pour tout $\ell \in \Omega_1^\star$, on impose que l'ensemble Ω_1^ℓ soit non vide : chaque "ℓ" associé a une grande valeur singulière de $X^\star$ est ciblée par au moins un "s" associé aux grandes valeurs singulières de X.

Ω_2^ℓ et Ω_3^ℓ peuvent être vides ou non : chaque $\ell \in \Omega_1^\star$ peut aussi être ciblé par quelques "s" appartenant à Ω_2 ou à Ω_3.

Application sur données simulées

Etude de la nouvelle proxy pour l'accuracy

Estimation des paramètres de nuisance avec un grand nombre de marqueurs chez Trainings $\Rightarrow$ nombre de marqueurs différents pour TESTS et Trainings

Contexte des simulations :

- Tailles des échantillons : 100 TESTS + 500 Trainings
- $T = 1$, 4 ou 6 Morgans
- Différentes densités de marqueurs
  - 1000 marqueurs pour Trainings et 500 marqueurs pour TESTS
  - 2000 marqueurs pour Trainings et 1000 marqueurs pour TESTS
- Différentes configurations de corrélation entre les régresseurs
  - 30, 50 ou 70 générations
- $\theta^\star$ vecteur de taille 25
  - 25 QTLs avec effet $+0.45$

Application sur données simulées

1000 marqueurs pour Trainings / 500 marqueurs pour TESTS

500 marqueurs pour Trainings / 500 marqueurs pour TESTS

T	Méthode	50 générations	70 générations	100 générations	$\overline{\text{MSE}}$
	Acc. Emp.	0.5287	0.5396	0.5173	
	$\hat\rho(\hat X^\star, \hat\theta^\star_{LASSO})$	0.4370 (0.0175)	0.4638 (0.0013)	0.4642 (0.0092)	0.0093
1	$\hat\rho(\hat X^\star, \hat\theta^\star_{GPLASSO})$	0.4033 (0.0239)	0.4469 (0.0163)	0.4471 (0.0115)	0.0172
	$\hat\rho(\hat X^\star, \hat\theta^\star_{ADLASSO})$	0.5371 (0.0073)	0.5691 (0.0063)	0.5589 (0.0069)	0.0068
	$\breve\rho^{pLD}(\hat\theta_{ADLASSO})$	0.5011 (0.0098)	0.5324 (0.0079)	0.5172 (0.0049)	0.0075
	$\hat\rho^{pLD}(\hat\theta_{ADLASSO})$	0.5411 (0.0099)	0.5758 (0.0094)	0.5690 (0.0087)	0.0093
	Acc. Emp.	0.3909	0.3772	0.3217	
	$\hat\rho(\hat X^\star, \hat\theta^\star_{LASSO})$	0.3397 (0.0112)	0.3436 (0.0132)	0.2629 (0.0146)	0.0130
4	$\hat\rho(\hat X^\star, \hat\theta^\star_{GPLASSO})$	0.2413 (0.0334)	0.3059 (0.0179)	0.2178 (0.0228)	0.0247
	$\hat\rho(\hat X^\star, \hat\theta^\star_{ADLASSO})$	0.4677 (0.01293)	0.4821 (0.0222)	0.4093 (0.0164)	0.0172
	$\breve\rho^{pLD}(\hat\theta_{ADLASSO})$	0.2599 (0.0389)	0.2647 (0.0355)	0.0846 (0.0722)	0.0489
	$\hat\rho^{pLD}(\hat\theta_{ADLASSO})$	0.2970 (0.0336)	0.3182 (0.0306)	0.0986 (0.0693)	0.0445
	Acc. Emp.	0.3749	0.3319	0.3155	
	$\hat\rho(\hat X^\star, \hat\theta^\star_{LASSO})$	0.37 (0.0034)	0.3548 (0.0094)	0.3415 (0.0093)	0.0074
6	$\hat\rho(\hat X^\star, \hat\theta^\star_{GPLASSO})$	0.3395 (0.01132)	0.3259 (0.0093)	0.3048 (0.0094)	0.0100
	$\hat\rho(\hat X^\star, \hat\theta^\star_{ADLASSO})$	0.5045 (0.02488)	0.4981 (0.0355)	0.4703 (0.0317)	0.0307
	$\breve\rho^{pLD}(\hat\theta_{ADLASSO})$	0.2351 (0.0436)	0.2383 (0.0358)	0.2423 (0.0307)	0.0367
	$\hat\rho^{pLD}(\hat\theta_{ADLASSO})$	0.1929 (0.0519)	0.1906 (0.0397)	0.2045 (0.0319)	0.0412

- proxy avec même nombre de marqueurs chez TRN et TESTS (DL parfait)
- nouvelle proxy, plus de marqueurs chez TRN que chez TESTS (DL imparfait)

performances nouvelle proxy $>$ performances ancienne proxy

Application sur données simulées

2000 marqueurs pour Trainings / 1000 marqueurs pour TESTS

1000 marqueurs pour Trainings / 1000 marqueurs pour TESTS

T	Méthode	50 générations	70 générations	100 générations	$\overline{\text{MSE}}$
	Emp. Acc.	0.5239	0.5561	0.5907	
	$\hat\rho(\hat X^\star, \hat\theta^\star_{LASSO})$	0.4218 (0.0181)	0.4213 (0.0224)	0.4676 (0.0220)	0.0208
1	$\hat\rho(\hat X^\star, \hat\theta^\star_{GPLASSO})$	0.3856 (0.0269)	0.3949 (0.0309)	0.4546 (0.0247)	0.0275
	$\hat\rho(\hat X^\star, \hat\theta^\star_{ADLASSO})$	0.5261 (0.0061)	0.5298 (0.0043)	0.5709 (0.0057)	0.0054
	$\check\rho^{pLD}(\hat\theta_{ADLASSO})$	0.4624 (0.0096)	0.4734 (0.0114)	0.5241 (0.0092)	0.0101
	$\hat\rho^{pLD}(\hat\theta_{ADLASSO})$	0.5107 (0.0068)	0.5153 (0.0062)	0.5641 (0.0065)	0.0065
	Emp. Acc.	0.4244	0.4027	0.4162	
	$\hat\rho(\hat X^\star, \hat\theta^\star_{LASSO})$	0.3614 (0.013)	0.3224 (0.0193)	0.3478 (0.0156)	0.0159
4	$\hat\rho(\hat X^\star, \hat\theta^\star_{GPLASSO})$	0.2974 (0.0260)	0.2521 (0.0403)	0.2929 (0.0256)	0.0306
	$\hat\rho(\hat X^\star, \hat\theta^\star_{ADLASSO})$	0.5063 (0.0147)	0.4642 (0.0146)	0.5001 (0.0152)	0.0148
	$\check\rho^{pLD}(\hat\theta_{ADLASSO})$	0.3037 (0.0291)	0.2441 (0.0414)	0.2906 (0.0328)	0.0344
	$\hat\rho^{pLD}(\hat\theta_{ADLASSO})$	0.3612 (0.0226)	0.3205 (0.0305)	0.3483 (0.0259)	0.0263
	Emp. Acc.	0.3724	0.4037	0.3477	
	$\hat\rho(\hat X^\star, \hat\theta^\star_{LASSO})$	0.3215 (0.0127)	0.3325 (0.0135)	0.2709 (0.0167)	0.0143
6	$\hat\rho(\hat X^\star, \hat\theta^\star_{GPLASSO})$	0.2619 (0.0236)	0.2799 (0.0240)	0.2071 (0.0299)	0.0258
	$\hat\rho(\hat X^\star, \hat\theta^\star_{ADLASSO})$	0.4863 (0.0212)	0.4966 (0.0144)	0.4401 (0.0167)	0.0174
	$\check\rho^{pLD}(\hat\theta_{ADLASSO})$	0.2024 (0.0478)	0.2309 (0.0499)	0.1844 (0.0413)	0.0463
	$\hat\rho^{pLD}(\hat\theta_{ADLASSO})$	0.2510 (0.0399)	0.2935 (0.0397)	0.2347 (0.0324)	0.0373

- proxy avec même nombre de marqueurs chez TRN et TESTS (DL parfait)
- nouvelle proxy, plus de marqueurs chez TRN que chez TESTS (DL imparfait)

performances nouvelle proxy $>$ performances ancienne proxy

Application sur données réelles de riz

Nombre de marqueurs nécessaires pour une prédiction précise des TESTS

Estimation des paramètres de nuisance avec un grand nombre de marqueurs chez Trainings $\Rightarrow$ nombre de marqueurs différents pour TESTS et Trainings

Date de floraison chez le riz (Spindel et al, Plos Genetics 2015)

- $K = 73{,}147$ pour les Trainings
- 4 densités de marqueurs pour les TESTS (448, 781, 1553 et 3076)
- 252 Trainings, 63 TESTS (i.e. 80% et 20%) + 100 tirages

Méthode	448 SNPs	781 SNPs	1553 SNPs	3076 SNPs	$\overline{\text{MSE}}$
Acc. Emp.	0.4789	0.4919	0.5275	0.5242	
$\hat{\rho}(\hat{X}^\star, \hat{\theta}^\star_{LASSO})$	0.4621 (0.0244)	0.4653 (0.0226)	0.4737 (0.0254)	0.4728 (0.0263)	0.0247
$\hat{\rho}(\hat{X}^\star, \hat{\theta}^\star_{ADLASSO})$	0.4269 (0.0355)	0.4379 (0.0376)	0.4520 (0.0419)	0.4461 (0.0430)	0.0395
$\hat{\rho}^{pLD}(\hat{\theta}_{ADLASSO})$	0.3662 (0.0454)	0.4202 (0.0281)	0.4919 (0.0215)	0.4952 (0.0342)	0.0323

- proxy avec même nombre de marqueurs chez TRN et TESTS (DL parfait)
- nouvelle proxy, plus de marqueurs chez TRN que chez TESTS (DL imparfait)

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 $\widetilde{\theta}$

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

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 :

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

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

...

Les 3 configurations possibles (résultat non asymptotique)

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

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

   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

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

   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

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

Références

- Rabier et Grusea (Journal of the Royal Statistical Society Series C, 2021). “Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium”
- Rabier, Mangin, Grusea (Scandinavian Journal of Statistics, 2019). “On the accuracy in high dimensional models and its application to genomic selection”
- Shao et Deng (Annals of statistics, 2012). “Estimation in high-dimensional linear models with deterministic design matrices”
- Fan et Lv (Journal of the Royal Statistical Society Series B, 2008). “Sure independence screening for ultrahigh dimensional feature space”.
- Spindel, Begum, ..., Jannink, McCouch (PLoS Genetics, 2015). “Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines”
- Ferrao, Ferrao, ..., Stephens, Garcia (Heredity, 2018). “Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models”.

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

Valeurs singulières

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

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

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

Paramètre de régularisation

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

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

on discute le lien entre les valeurs singulières $d_s^\star$, d_s, λ
et projection du signal sur sous espaces associés

Liens valeurs singulières / paramètre de régularisation

- Considérons les partitions suivantes $\Omega_1^\star$, $\Omega_2^\star$, $\Omega_3^\star$ et Ω_1, Ω_2, Ω_3 :

$$\Omega_1^\star := \left\{ \ell \,\middle|\, \lambda := o(d_\ell^{\star 2}) \right\}, \quad \Omega_1 := \left\{ s \,\middle|\, \lambda = o(d_s^2) \right\}$$

$$\Omega_2^\star := \left\{ \ell \,\middle|\, d_\ell^{\star 2} \sim \frac{1}{C_\ell^\star} \lambda \text{ with } C_\ell^\star > 0 \right\}, \quad \Omega_2 := \left\{ s \,\middle|\, d_s^2 \sim \frac{1}{C_s} \lambda \text{ avec } C_s > 0 \right\}$$

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

Quelques conditions

Pour chaque $\ell \in \{1, \ldots, r^\star\}$, prenant en compte que $\left\| U^{\star(\ell)} \right\|^2 = 1$, on definit $\xi_k^{(\ell)} \in]0, 1]$, $k = 1, 2, 3$ par :

$$(C0^\star) \text{ Si } \#\Omega_k^\ell \neq 0 \text{ , } \left\| U^{(s)} U^{(s)\prime} U^{\star(\ell)} \right\|^2 \sim \frac{\xi_k^{(\ell)}}{\#\Omega_k^\ell} \;\; \forall s \in \Omega_k^\ell \text{ ,}$$

avec $\sum_{k | \Omega_k^\ell \neq \emptyset} \xi_k^{(\ell)} \leq 1$.

Quelques conditions

- $(C1^\star)$ $\dfrac{n^{2\tau}}{r^\star} \displaystyle\sum_{\ell \in \Omega_1^\star} d_\ell^{\star 2} \to +\infty$
- $(C2)$ $\displaystyle\sum_{s \in \Omega_3} d_s^2 = o(\lambda)$
- $(C3)$ $\displaystyle\sum_{s \in \Omega_3} d_s^4 = o(\lambda^2)$
- $(C4^\star)$ $\dfrac{n^{2\tau}}{r^\star} = o(1/\lambda)$
- $(C5)$ $\#\Omega_1 = O(1)$
- $(C6)$ $\#\Omega_2 = O(1)$
- $(C7^\star)$ $\dfrac{n^{2\tau}}{r^\star} \displaystyle\sum_{\ell \in \Omega_1^\star} \xi_2^{(\ell)} d_\ell^{\star 2} = o(1)$
- $(C8^\star)$ $\dfrac{n^{2\tau}}{r^\star} \displaystyle\sum_{\ell \in \Omega_1^\star} \xi_3^{(\ell)} d_\ell^{\star 2} = o(1)$

$(C1^\star)$, $(C4^\star)$, $(C7^\star)$, $(C8^\star)$ sont spécifiques à cette étude

$(C2)$, $(C3)$, $(C5)$, $(C6)$ étaient déjà présentes dans l'étude du LD parfait

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

Lemma (Convergence vers l'accuracy oracle)

Supposons que pour $k = 1, 2, 3$, nous avons $\Omega_k^{\ell} \cap \Omega_k^{\ell'} = \emptyset \ \forall \ell \neq \ell'$. De plus, supposons que le signal est projeté uniformément sur chaque espace Vect$\left\{ \boldsymbol{W}^{\star(\ell)} \right\}$, i.e.

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

De plus, $\forall \ell \in \Omega_1^{\star}$, assumons que $\Omega_1^{\ell} \neq \emptyset$ et que $\xi_1^{(\ell)} = \xi(n)$ avec $0 < b < \xi(n) \leq 1$.
Alors, en supposant les conditions $(C0^{\star} - C1^{\star} - C2 - C3 - C4^{\star} - C5 - C6 - C7^{\star} - C8^{\star})$,

- *lorsque n est grand, nous avons $\hat{\rho}_g \sim \sqrt{\xi(n)}\ \rho_g^{oracle}$*
- *Si $\forall \ell \in \Omega_1^{\star}$, $\xi_2^{(\ell)} = 1/n^{\theta_1}$ et $\xi_3^{(\ell)} = 1/n^{\theta_2}$ avec $\theta_1 > \psi$ et $\theta_2 > \psi$, alors nous avons $\hat{\rho}_g \longrightarrow \rho_g^{oracle}$.*