Sélection d'effets fixes dans les modèles linéaires mixtes de grande dimension

Florian ROHART, Magali SAN CRISTOBAL et Béatrice LAURENT

INRA Toulouse
INSA Toulouse et IMT

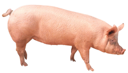
27 novembre 2014

Plan

Plan

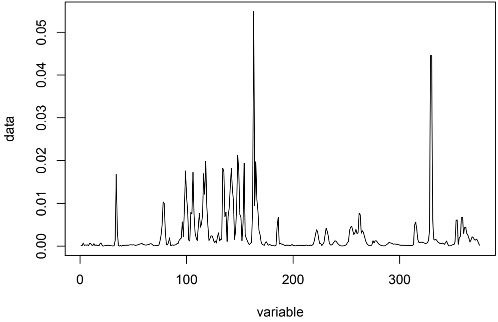
Motivation : prédire des phénotypes de production à l'aide de données métabolomiques


(a) Landrace


(b) Large White


(c) Pietrain

$n = 506$ individus, $p = 375$ variables métabolomiques, $\times 3$ si interactions avec la race



Premières analyses

Objectif double :

- Prédire un phénotype. Estimer le pouvoir prédictif du métabolome pour les phénotypes.
  Méthodes : PLS, sPLS, Random Forest,. . .
- Identifier les métabolites qui expliquent le plus un phénotype : **sélection de variables**.
  Méthodes : Lasso et ses extensions [Rohart et al 2012]

Premières analyses

Objectif double :

- Prédire un phénotype. Estimer le pouvoir prédictif du métabolome pour les phénotypes.
  Méthodes : PLS, sPLS, Random Forest,...
- Identifier les métabolites qui expliquent le plus un phénotype : **sélection de variables**.
  Méthodes : Lasso et ses extensions [Rohart et al 2012]

Dans cette première analyse, Rohart et al (2012) s'étaient placés dans le modèle linéaire :

$$Y = X\beta + \epsilon \text{ avec } \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

Y phénotype observé sur n individus,
$X = (X_1, \ldots, X_p)$ variables métabolomiques,
β vecteur de p paramètres inconnus.

Premières analyses

Objectif double :

- Prédire un phénotype. Estimer le pouvoir prédictif du métabolome pour les phénotypes.
  Méthodes : PLS, sPLS, Random Forest,. . .
- Identifier les métabolites qui expliquent le plus un phénotype : **sélection de variables**.
  Méthodes : Lasso et ses extensions [Rohart et al 2012]

Dans cette première analyse, Rohart et al (2012) s'étaient placés dans le modèle linéaire :

$$Y = X\beta + \epsilon \text{ avec } \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

Y phénotype observé sur n individus,
$X = (X_1, \ldots, X_p)$ variables métabolomiques,
β vecteur de p paramètres inconnus.

Or, les données sont structurées en familles de demi-frères de pères, et en lots d'animaux (effets aléatoires).

Cadre

Sélection de variables en modèle linéaire mixte gaussien

Plan

Plan

Sélection de variables en modèle linéaire mixte gaussien de grande dimension

Le modèle :

$$y = X\beta + \sum_{k=1}^{q} Z_k u_k + e \text{ avec } u_k \sim \mathcal{N}(0, \sigma_k^2 I_{N_k}), e \sim \mathcal{N}(0, \sigma_e^2 I_n) \quad (1)$$

Sélection de variables en modèle linéaire mixte gaussien de grande dimension

Le modèle :

$$y = X\beta + \sum_{k=1}^{q} Z_k u_k + e \text{ avec } u_k \sim \mathcal{N}(0, \sigma_k^2 I_{N_k}), e \sim \mathcal{N}(0, \sigma_e^2 I_n) \quad (1)$$

Schelldorfer et al (2011) se placent dans le modèle marginal $y = X\beta + \epsilon$ avec $Var(\epsilon) = V$.

- Sélection de variables (β) avec LASSO (pénalisation ℓ_1) : package R lmmLasso.
- Même structure de groupe pour tous les Z_k.
- Résultats théoriques : consistence quand le nombre de groupes $\to \infty$.
- Temps de calcul prohibitif si grande taille d'échantillon (inversion matrice V).

Sélection de variables en modèle linéaire mixte gaussien de grande dimension

Le modèle :

$$y = X\beta + \sum_{k=1}^{q} Z_k u_k + e \text{ avec } u_k \sim \mathcal{N}(0, \sigma_k^2 I_{N_k}), e \sim \mathcal{N}(0, \sigma_e^2 I_n) \quad (1)$$

Schelldorfer et al (2011) se placent dans le modèle marginal $y = X\beta + \epsilon$ avec $Var(\epsilon) = V$.

- Sélection de variables (β) avec LASSO (pénalisation ℓ_1) : package R lmmLasso.
- Même structure de groupe pour tous les Z_k.
- Résultats théoriques : consistence quand le nombre de groupes $\rightarrow \infty$.
- Temps de calcul prohibitif si grande taille d'échantillon (inversion matrice V).

Bondell et al (2010) et Ibrahim et al (2011) ont choisi le modèle (1), la même structure des effets aléatoires

- Sélection d'effets fixes et aléatoires,
- Une re-paramétrisation des effets aléatoires (type Choleski) qui dépend de l'ordre dans lequel ils sont considérés (Muller et al 2013).

Notre approche

$\Rightarrow$ On va se placer dans le modèle (1), avec des structures des Z_k identiques ou pas, sans re-paramétrisation, et proposer un algorithme ECM (convergence vers un optimum local assurée).

Notre approche

⇒ On va se placer dans le modèle (1), avec des structures des Z_k identiques ou pas, sans re-paramétrisation, et proposer un algorithme ECM (convergence vers un optimum local assurée).

Remarque : Groll et Tutz (2014) ont choisi cette approche dans le cadre du modèle linéaire généralisé

Notre approche

$\Rightarrow$ On va se placer dans le modèle (1), avec des structures des Z_k identiques ou pas, sans re-paramétrisation, et proposer un algorithme ECM (convergence vers un optimum local assurée).

Remarque : Groll et Tutz (2014) ont choisi cette approche dans le cadre du modèle linéaire généralisé

$J = \{j, \beta_j \neq 0\}$. On suppose que $\sum_{k=1}^{q} N_k + |J| < n$.

Objectif

Estimer J, β, $\sigma_1^2, \ldots, \sigma_q^2$ et σ_e^2.

Fonction objectif

On considère $\{u_k\}_{k\in\mathcal{K}}$ comme des variables manquantes.
$\Phi = (\beta, \sigma_1^2, \dots, \sigma_q^2, \sigma_e^2)$ est le vecteur de paramètres à estimer
La log-vraisemblance des données complétées $x = (y, u)$ est

Log-vraisemblance

$$L(\Phi; x) = L_0(\beta, \sigma_e^2, \sigma_1^2, \dots, \sigma_q^2; \epsilon) + \sum_{k=1}^{q} L_k(\sigma_k^2; u_k) + C, \quad (2)$$

avec

$$-2L_0(\beta, \sigma_e^2, \sigma_u^2; \epsilon) = n\ln(\sigma_e^2) + \left|\left|y - X\beta - \sum Z_k u_k\right|\right|^2 / \sigma_e^2 \quad (3a)$$

$$-2L_k(\sigma_k^2; u_k) = N_k \ln(\sigma_k^2) + ||u_k||^2 / \sigma_k^2 \quad (3b)$$

Fonction objectif

On considère $\{u_k\}_{k\in\mathcal{K}}$ comme des variables manquantes.
$\Phi = (\beta, \sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2)$ est le vecteur de paramètres à estimer
La log-vraisemblance des données complétées $x = (y, u)$ est

Log-vraisemblance

$$L(\Phi; x) = L_0(\beta, \sigma_e^2, \sigma_1^2, \ldots, \sigma_q^2; \epsilon) + \sum_{k=1}^{q} L_k(\sigma_k^2; u_k) + C, \tag{2}$$

Fonction objectif

$$g(\Phi; x) = -2L(\Phi; x) + \lambda|\beta|_1 \tag{3}$$

Plan

Algorithme multicycle ECM

Initialisation :
$\mathcal{K} = \{1, \ldots, q\}$.
Initialisation des paramètres $(\sigma_{\mathcal{K}}^{2[0]}, \sigma_e^{2[0]}, \beta^{[0]})$.
Jusqu'à convergence :

1. $u^{[t+1/2]} = E(u|y, reste^{[t]}) = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t]})$
2. $\beta^{[t+1]} = \underset{\beta}{Argmin} \left(\left|\left| (y - Zu^{[t+1/2]}) - X\beta \right|\right|^2 + \lambda\sigma_e^{2[t]} |\beta|_1 \right)$
3. $u^{[t+1]} = E(u|y, reste^{[t+1]})$
4.
   (a) Pour k dans $\mathcal{K}$
   $\sigma_k^{2[t+1]} = E(u_k' u_k | y, reste^{[t+1]})/N_k$
   $\sigma_{k,\ell}^{[t+1]} = E(u_k' u_\ell | y, reste^{[t+1]})/N_k$
   si $\left(\left|\left| [u^{[t+1]}]_k \right|\right|^2 / N_k < 10^{-6} \right)$ alors $\mathcal{K} = \mathcal{K}_{-k}$
   (b) $\sigma_e^{2(t+1)} = E(e'e|y, reste^{[t+1]})/n$

Algorithme multicycle ECM

Initialisation :
$\mathcal{K} = \{1, \dots, q\}$.
Initialisation des paramètres $(\sigma_{\mathcal{K}}^{2[0]}, \sigma_e^{2[0]}, \beta^{[0]})$.
Jusqu'à convergence :

1. $u^{[t+1/2]} = E(u|y, reste^{[t]})$
2. Sélection de variables et estimation de β dans le modèle linéaire $Y - Zu^{[t+1/2]} = X\beta + \epsilon^{[t]}$, où $\epsilon^{[t]}$ est un bruit i.i.d gaussien.
3. $u^{[t+1]} = E(u|y, reste^{[t+1]})$
4.
(a) Pour k dans $\mathcal{K}$
$\sigma_k^{2\,[t+1]} = E(u_k' u_k|y, reste^{[t+1]})/N_k$
$\sigma_{k,\ell}{}^{[t+1]} = E(u_k' u_\ell|y, reste^{[t+1]})/N_k$
si $\left(||[u^{[t+1]}]_k||^2/N_k < 10^{-6}\right)$ alors $\mathcal{K} = \mathcal{K}_{-k}$
(b) $\sigma_e^{2(t+1)} = E(e'e|y, reste^{[t+1]})/n$

Plan

Cadre des simulations

$n = 120$
$q = 2$ effets aléatoires ($\sigma_1^2 = \sigma_2^2 = 1$)
$J = \{1, 2, 3, 4, 5\}$ et $\beta_J = 3/4$
Signal / bruit de l'ordre de 0.9 (difficile)

- M_1 : $p = 80$. Petite dimension. Effets aléatoires indépendants. Ce modèle est ajusté en supposant à tort $q = 3$.
- M_2 : $p = 300$. Grande dimension. Les 2 effets aléatoires sont corrélés ($\rho = 0.5$).
- M_3 : $p = 600$. Très grande dimension. Effets aléatoires indépendants.
- M_4 : $p = 600$. Très grande dimension. Structures différentes pour les effets aléatoires (exit lmmlasso).

Cadre des simulations

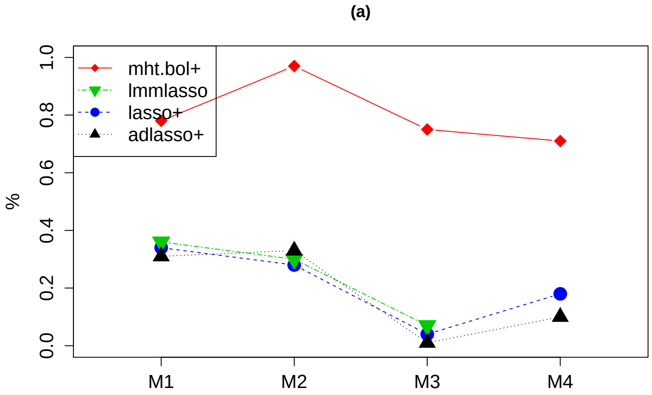
$n = 120$
$q = 2$ effets aléatoires ($\sigma_1^2 = \sigma_2^2 = 1$)
$J = \{1, 2, 3, 4, 5\}$ et $\beta_J = 3/4$
Signal / bruit de l'ordre de 0.9 (difficile)

- M_1 : $p = 80$. Petite dimension. Effets aléatoires indépendants. Ce modèle est ajusté en supposant à tort $q = 3$.
- M_2 : $p = 300$. Grande dimension. Les 2 effets aléatoires sont corrélés ($\rho = 0.5$).
- M_3 : $p = 600$. Très grande dimension. Effets aléatoires indépendants.
- M_4 : $p = 600$. Très grande dimension. Structures différentes pour les effets aléatoires (exit lmmlasso).

Méthodes :

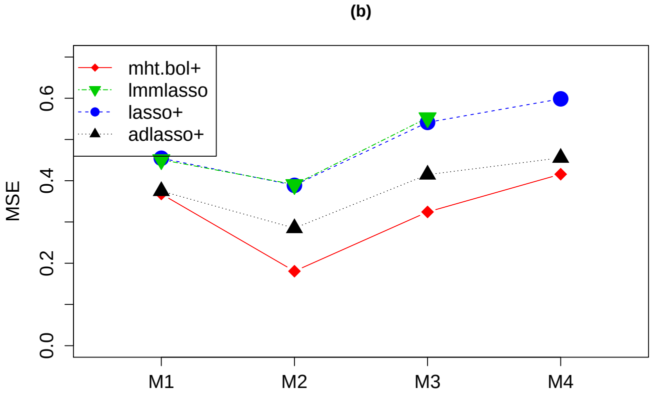
- Modèles linéaires : Lasso (Tibshirani 1996), adaptive Lasso : adLasso (Zou 2006), mht.bol (Rohart 2012)
- Modèles linéaires mixtes : Lasso+, adLasso+, mht.bol+, et lmmLasso (Schelldorfer et al 2011)

mht.bol+ surpasse les autres, en terme de bons modèles sélectionnés



Pourcentage de bons modèles sélectionnés
Trop d'effets fixes en général

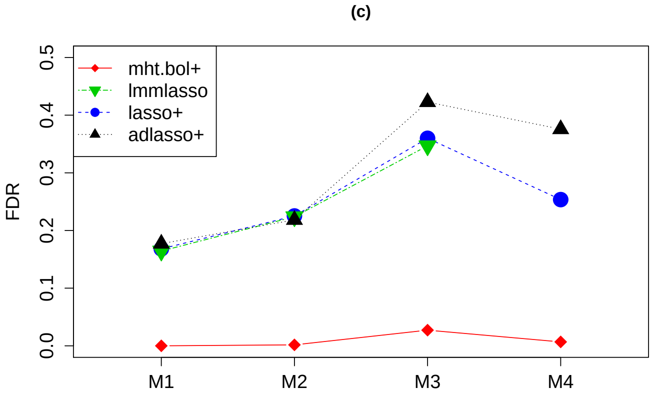
mht.bol+ surpasse les autres, en terme de MSE



Erreur quadratique moyenne $||X\beta - X\hat{\beta}||_n^2$

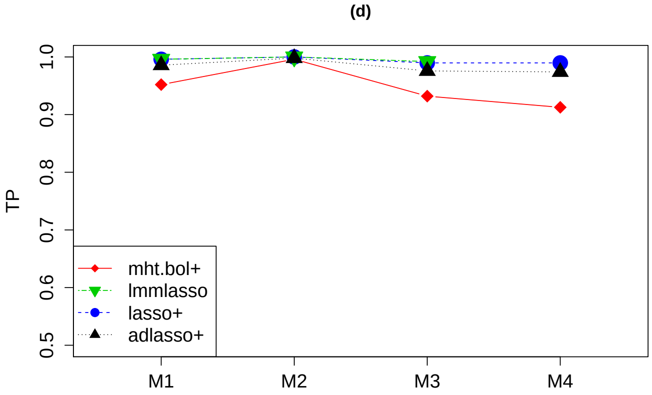
Remarque : lmmlasso = lasso+ (même modèle, algorithme différent)

mht.bol+ surpasse les autres, en terme de FDR



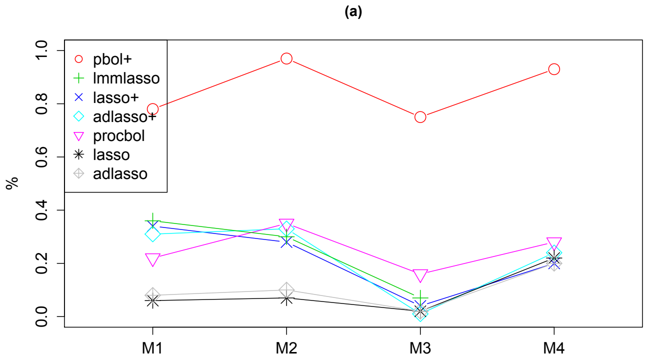
Taux de faux positifs

mht.bol+ un peu moins bon, en terme de vrais positifs



Taux de vrais positifs
C'est le (petit) prix à payer ...

On perd beaucoup en ignorant les effets aléatoires



Pourcentage de bons modèles sélectionnés
Dans tous les cas, il est vraiment préférable de ne pas oublier la structure due aux facteurs à effets aléatoires

Plan

Application aux données réelles : réduction de la variance résiduelle et du nombre de variables sélectionnées

$$y = X_{race}\beta_{race} + X_{metab}\beta_{metab} + Z_{lot}u_{lot} + Z_{fam}u_{fam} + e, \tag{4}$$

où y est un phénotype (Consommation Moyenne Journalière).
Pas de sélection sur β_{race}.

Application aux données réelles : réduction de la variance résiduelle et du nombre de variables sélectionnées

$$y = X_{race}\beta_{race} + X_{metab}\beta_{metab} + Z_{lot}u_{lot} + Z_{fam}u_{fam} + e, \quad (4)$$

où y est un phénotype (Consommation Moyenne Journalière).
Pas de sélection sur β_{race}.

	$\lvert\hat{J}\rvert$	$\hat{\sigma}_e^2(\times 10^{-2})$	$\hat{\sigma}_{lot}^2(\times 10^{-3})$	$\hat{\sigma}_{fam}^2(\times 10^{-3})$
Lasso	14	3.8	-	-
adLasso	21	3.4	-	-
mht.bol	11	4.1	-	-
Lasso+	11	3.2	3.2	6.4
adLasso+	10	3.3	2.5	6.5
mht.bol+	5	3.4	5.9	6.5

- La prise en compte d'effets aléatoires structurant les données réduit la variance résiduelle et le nombre de variables sélectionnées.

Application aux données réelles : réduction de la variance résiduelle et du nombre de variables sélectionnées

$$y = X_{race}\beta_{race} + X_{metab}\beta_{metab} + Z_{lot}u_{lot} + Z_{fam}u_{fam} + e, \quad (4)$$

où y est un phénotype (Consommation Moyenne Journalière).
Pas de sélection sur β_{race}.

	$\lvert\hat{J}\rvert$	$\hat{\sigma}_e^2(\times 10^{-2})$	$\hat{\sigma}_{lot}^2(\times 10^{-3})$	$\hat{\sigma}_{fam}^2(\times 10^{-3})$
Lasso	14	3.8	-	-
adLasso	21	3.4	-	-
mht.bol	11	4.1	-	-
Lasso+	11	3.2	3.2	6.4
adLasso+	10	3.3	2.5	6.5
mht.bol+	5	3.4	5.9	6.5

- La prise en compte d'effets aléatoires structurant les données réduit la variance résiduelle et le nombre de variables sélectionnées.
- La meilleure méthode (mht.bol+) au sens des simulations sélectionne 5 variables seulement.

Application aux données réelles : réduction de la variance résiduelle et du nombre de variables sélectionnées

$$y = X_{race}\beta_{race} + X_{metab}\beta_{metab} + Z_{lot}u_{lot} + Z_{fam}u_{fam} + e, \qquad (4)$$

où y est un phénotype (Consommation Moyenne Journalière).
Pas de sélection sur β_{race}.

	$\lvert\hat{J}\rvert$	$\hat{\sigma}_e^2(\times 10^{-2})$	$\hat{\sigma}_{lot}^2(\times 10^{-3})$	$\hat{\sigma}_{fam}^2(\times 10^{-3})$
Lasso	14	3.8	-	-
adLasso	21	3.4	-	-
mht.bol	11	4.1	-	-
Lasso+	11	3.2	3.2	6.4
adLasso+	10	3.3	2.5	6.5
mht.bol+	5	3.4	5.9	6.5

- La prise en compte d'effets aléatoires structurant les données réduit la variance résiduelle et le nombre de variables sélectionnées.
- La meilleure méthode (mht.bol+) au sens des simulations sélectionne 5 variables seulement.
- Cohérence biologique : la créatinine (liée à la masse musculaire) est toujours sélectionnée.

Amélioration du temps de calcul sur les données réelles

Un seul effet aléatoire afin de pouvoir utiliser le package lmmLasso.

Méthodes	temps CPU en secondes
lmmLasso	63
Lasso+ (package MMS)	1

Plan

Conclusion

- La sélection de variables (effets fixes) dans un modèle linéaire mixte gaussien est possible en temps raisonnable sur de gros fichiers et en grande dimension ($n < p$).

Conclusion

- La sélection de variables (effets fixes) dans un modèle linéaire mixte gaussien est possible en temps raisonnable sur de gros fichiers et en grande dimension ($n < p$).
- Le modèle peut inclure plusieurs effets aléatoires, de structure différentes ou non, et dans ce cas corrélés ou non.

Conclusion

- La sélection de variables (effets fixes) dans un modèle linéaire mixte gaussien est possible en temps raisonnable sur de gros fichiers et en grande dimension ($n < p$).
- Le modèle peut inclure plusieurs effets aléatoires, de structure différentes ou non, et dans ce cas corrélés ou non.
- Selon nos simulations, la méthode de sélection de variables la plus performante est celle de Rohart (2012) basée sur les tests multiples.

Références

- Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection of fixed and random effects in linear mixed-effects models. Biometrics, 66 :1069-1077.
- Groll, A., Tutz, G. (2014). Variable selection for generalized linear mixed models by 1-penalized estimation. Stat. Comput. 24, 137–154.
- Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and Random Effects Selection in Mixed Effects Models. Biometrics, 67 :495-503.
- Müller, S., Scealy, J., and Welsch, A. (2013). Model selection in linear mixed model. Statistical Science, to appear.
- Rohart, F. (2012). Multiple Hypotheses Testing For Variable Selection. http ://arxiv.org/abs/1106.3415
- Rohart, F., et al. (2012). Phenotypic Prediction Based on Metabolomic Data on the Growing Pig from three main European Breeds. Journal of Animal Science, 90 :4729-40.
- Rohart F, San Cristobal M, Laurent B (2014) Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. Computational Statistics and Data Analysis 80, 209-222.
- Schelldorfer, J., Buhlmann, P., and van de Geer, S. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ1-Penalization. Scandinavian Journal of Statistics, 38 :197-214.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, B 58(1) :267-288.
- Zou, H. (2006). The adaptive lasso and its oracle properties. J. Amer. Statist.