Variable selection in high-dimensional genetic data

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
Department of Diagnostic Radiology
McGill University

sahirbhatnagar.com

Feb. 23, 2021

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

Other applications

Extensions

- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

Setting

- We are concerned with the analysis of data in which we are attempting to predict an outcome Y using a number of explanatory factors X_1, X_2, X_3, . . ., some of which may not be particularly useful

Setting

- We are concerned with the analysis of data in which we are attempting to predict an outcome Y using a number of explanatory factors $X_1, X_2, X_3, \ldots$, some of which may not be particularly useful
- Although the methods we will discuss can be used solely for prediction (i.e., as a "black box"), I will adopt the perspective that we would like the statistical methods to be interpretable and to explain something about the relationship between the X and Y

Setting

- We are concerned with the analysis of data in which we are attempting to predict an outcome Y using a number of explanatory factors X_1, X_2, X_3, . . ., some of which may not be particularly useful
- Although the methods we will discuss can be used solely for prediction (i.e., as a "black box"), I will adopt the perspective that we would like the statistical methods to be interpretable and to explain something about the relationship between the X and Y
- Regression models are an attractive framework for approaching problems of this type, and the focus today will be on extending classical regression modeling to deal with high-dimensional data

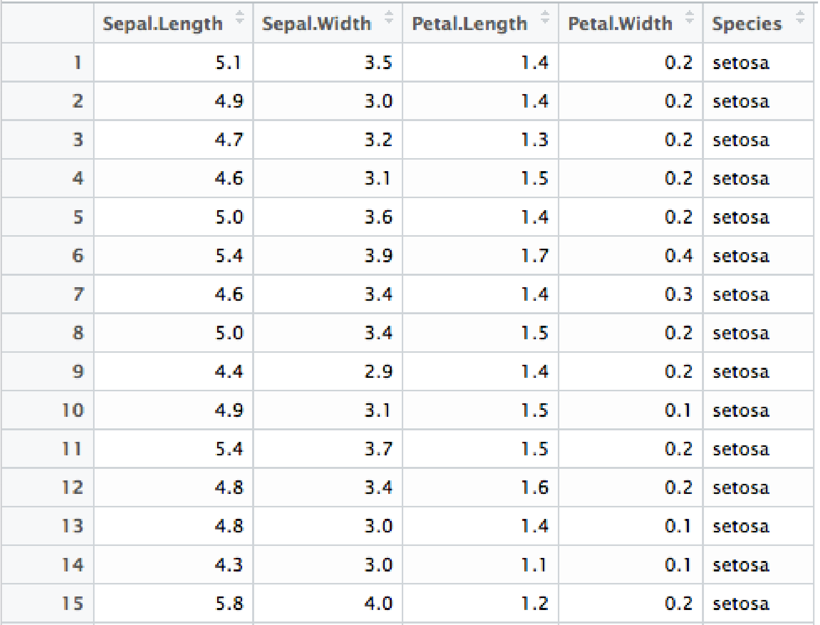
Classical Methods

- A nice and powerful toolbox for analyzing the more traditional datasets where the sample size (N) is far **greater than** the number of covariates (p):
  - linear regression, logistic regression, LDA, QDA, glm,
  - regression spline, smoothing spline, kernel smoothing, local smoothing, GAM,
  - Neural Network, SVM, Boosting, Random Forest, ...

Classical Methods

- A nice and powerful toolbox for analyzing the more traditional datasets where the sample size (N) is far **greater than** the number of covariates (p):
  - linear regression, logistic regression, LDA, QDA, glm,
  - regression spline, smoothing spline, kernel smoothing, local smoothing, GAM,
  - Neural Network, SVM, Boosting, Random Forest, ...

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{12} & \cdots & x_{1p} \\ x_{31} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & x_{np} \end{bmatrix}$$



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa

Classical Linear Regression

Data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ iid from

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$$

where $E(\epsilon|\mathbf{x}) = 0$, and $\dim(x) = p$. To include an intercept, we can set $\mathbf{x}_1 \equiv 1$. Using Matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

The least squares estimator

$$\widehat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

$$\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Classical Linear Regression

Data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ iid from

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$$

where $E(\epsilon|\mathbf{x}) = 0$, and $\dim(x) = p$. To include an intercept, we can set $\mathbf{x}_1 \equiv 1$. Using Matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

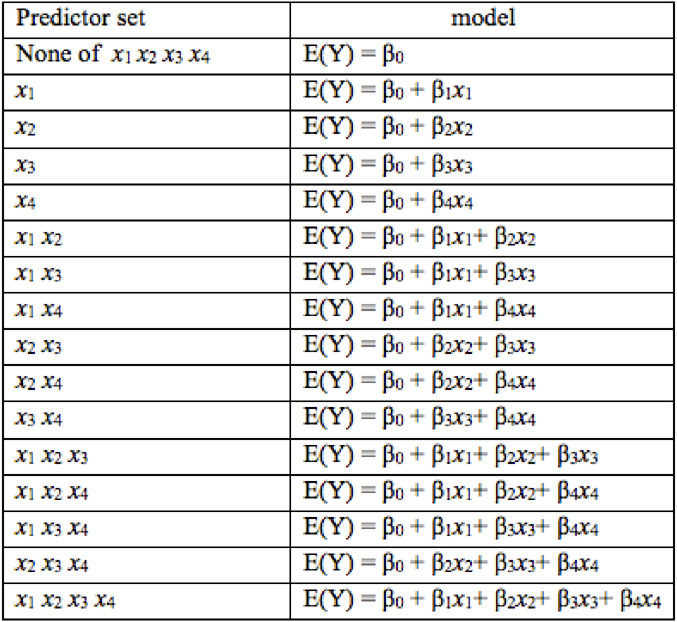
The least squares estimator

$$\widehat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

$$\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- **Question:** How to find the important variables $\mathbf{x}_j$?

Best-subset Selection (Beal et al. 1967, Biometrika)



Predictor set	model
None of $x_1\ x_2\ x_3\ x_4$	$E(Y) = \beta_0$
x_1	$E(Y) = \beta_0 + \beta_1 x_1$
x_2	$E(Y) = \beta_0 + \beta_2 x_2$
x_3	$E(Y) = \beta_0 + \beta_3 x_3$
x_4	$E(Y) = \beta_0 + \beta_4 x_4$
$x_1\ x_2$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
$x_1\ x_3$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$
$x_1\ x_4$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_4 x_4$
$x_2\ x_3$	$E(Y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$
$x_2\ x_4$	$E(Y) = \beta_0 + \beta_2 x_2 + \beta_4 x_4$
$x_3\ x_4$	$E(Y) = \beta_0 + \beta_3 x_3 + \beta_4 x_4$
$x_1\ x_2\ x_3$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
$x_1\ x_2\ x_4$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$
$x_1\ x_3\ x_4$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4$
$x_2\ x_3\ x_4$	$E(Y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
$x_1\ x_2\ x_3\ x_4$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

Which variables are important?

- Scientists know only a small subset of variables (such as genes) are important for the response variable.
- An old Idea: try all possible subset models and pick the best one.
- Fit a subset of predictors to the linear regression model. Let S be the subset predictors, e.g., S = {1, 3, 7}.

$$C_p = \frac{\mathrm{RSS}_S}{\sigma^2} - (n - 2|S|) = \frac{\mathrm{RSS}_S}{\sigma^2} + 2|S| - n$$

- We pick the model with the smallest C_p value.

Model selection criteria

Minimizing C_p is equivalent to minimizing

$$\|\underline{\mathbf{y}} - \mathbf{X}_S\widehat{\boldsymbol{\beta}}_S\|^2 + 2|S|\sigma^2.$$

which is AIC score.
Many popular model selection criteria can be written as

$$\|\underline{\mathbf{y}} - \mathbf{X}_S\widehat{\boldsymbol{\beta}}_S\|^2 + \lambda|S|\sigma^2.$$

- BIC uses $\lambda = \sigma\sqrt{\log(n)/n}$.

Remarks

Best subset selection plus model selection criteria (AIC, BIC, etc.)

- Computing all possible subset models is a combinatorial optimization problem (NP hard)
- Instability in the selection process (Breiman, 1996)

Ridge Regression (Hoerl & Kennard 1970, Technometrics)

- $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_2^2$
- $||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2$

Ridge Regression (Hoerl & Kennard 1970, Technometrics)

- $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_2^2$
- $||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2$
- $\widehat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \rightarrow$ exact solution
- $\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

Ridge Regression (Hoerl & Kennard 1970, Technometrics)

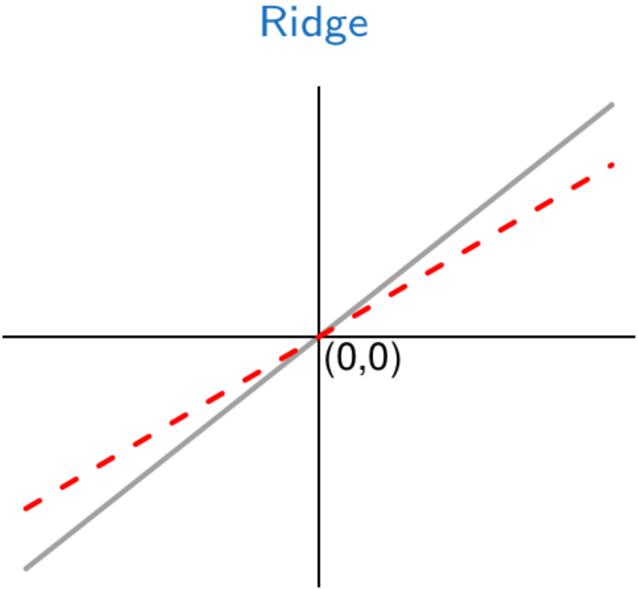
- $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\boldsymbol{\beta}||_2^2$
- $||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2$
- $\widehat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \rightarrow$ exact solution
- $\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
- Let $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_{p\times p}$

Ridge Regression (Hoerl & Kennard 1970, Technometrics)

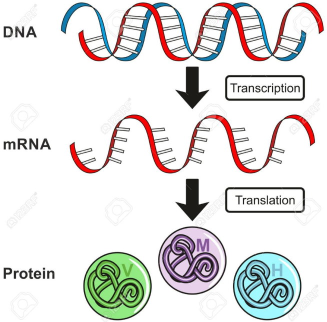
- $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_2^2$
- $||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2$
- $\widehat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \rightarrow$ exact solution
- $\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
- Let $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_{p\times p}$

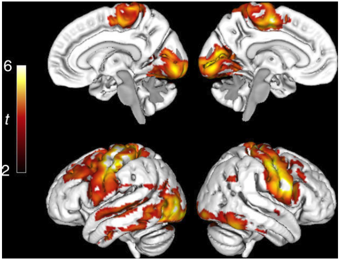
$$\hat{\beta}_{j_{(Ridge)}} = \frac{\hat{\beta}_{j_{(MCO)}}}{1+\lambda}$$

Least squares vs. Ridge



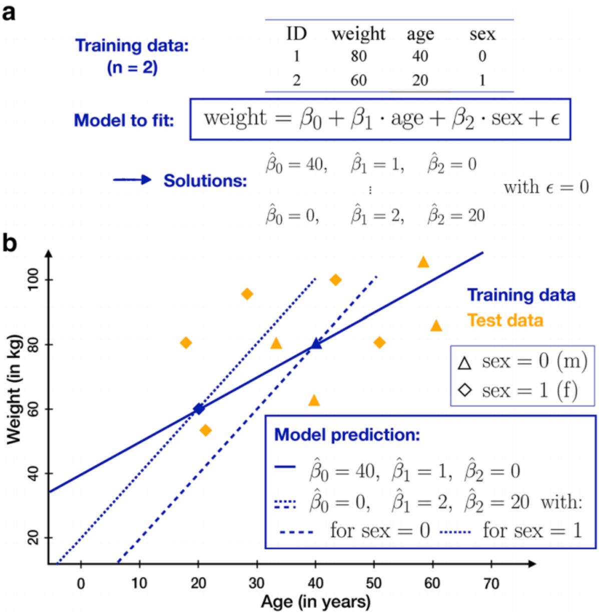
High-dimensional data ($n << p$)





$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{np} \end{bmatrix}$$

Why can’t we fit OLS to High-dimensional data?



a

Training data: (n = 2)

ID	weight	age	sex
1	80	40	0
2	60	20	1

Model to fit: $\boxed{\text{weight} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sex} + \epsilon}$

→ Solutions:

$$\hat\beta_0 = 40, \quad \hat\beta_1 = 1, \quad \hat\beta_2 = 0$$
$$\vdots$$
$$\hat\beta_0 = 0, \quad \hat\beta_1 = 2, \quad \hat\beta_2 = 20$$

with $\epsilon = 0$

b

Model prediction:

— $\hat\beta_0 = 40, \ \hat\beta_1 = 1, \ \hat\beta_2 = 0$

$\hat\beta_0 = 0, \ \hat\beta_1 = 2, \ \hat\beta_2 = 20$ with:

- - - - for sex = 0 ······· for sex = 1

High-dimensional data ($n << p$)

- We will let
  - n denote the number of independent sampling units (e.g., number of patients)
  - p denote the number of features recorded for each unit

High-dimensional data ($n << p$)

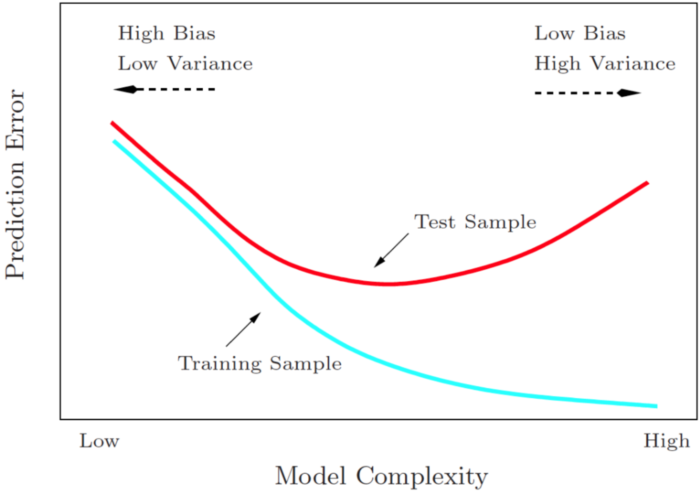
- We will let
  - n denote the number of independent sampling units (e.g., number of patients)
  - p denote the number of features recorded for each unit

- In high-dimensional data, p is large with respect to n
  - This certainly includes the case where $p > n$

High-dimensional data ($n << p$)

- We will let
  - n denote the number of independent sampling units (e.g., number of patients)
  - p denote the number of features recorded for each unit
- In high-dimensional data, p is large with respect to n
  - This certainly includes the case where $p > n$
  - However, the ideas we discuss in this course are also relevant to many situations in which $p < n$; for example, if $n = 100$ and $p = 80$, we probably don't want to use ordinary least squares

A fundamental picture for data science



ESL, Hastie et al. 2009

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

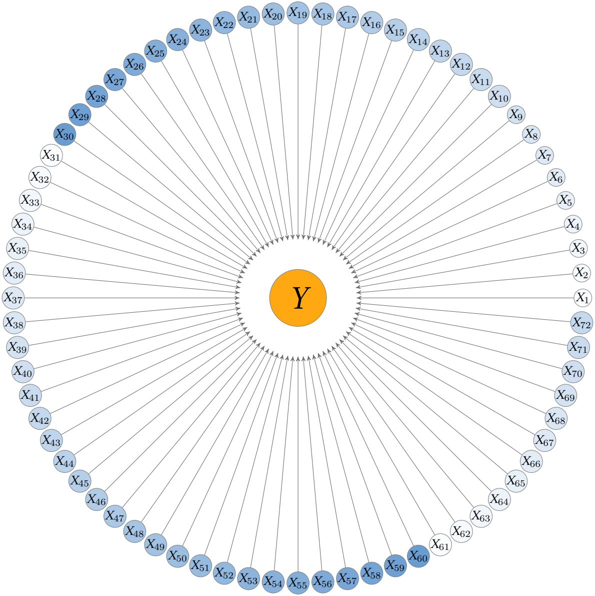
Other applications

Extensions

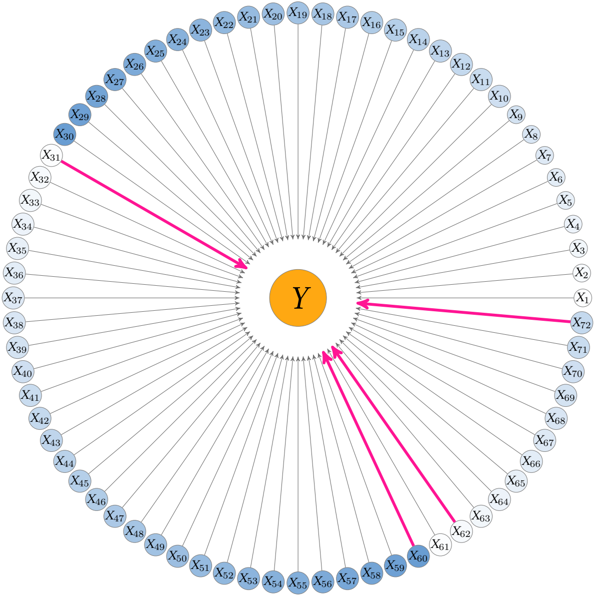
- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

Bet on Sparsity Principle



Bet on Sparsity Principle



Bet on Sparsity Principle

Use a procedure that does well in sparse problems, since no procedure does well in dense problems.[1]

[1] The elements of statistical learning. Springer series in statistics, 2001.

Bet on Sparsity Principle

Use a procedure that does well in sparse problems, since no procedure does well in dense problems.[1]

- We often don't have enough data to estimate so many parameters
- Even when we do, we might want to identify a **relatively small number of predictors** ($k < N$) that play an important role
- Faster computation, easier to understand, and stable predictions on new datasets.

[1] The elements of statistical learning. Springer series in statistics, 2001.

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

Other applications

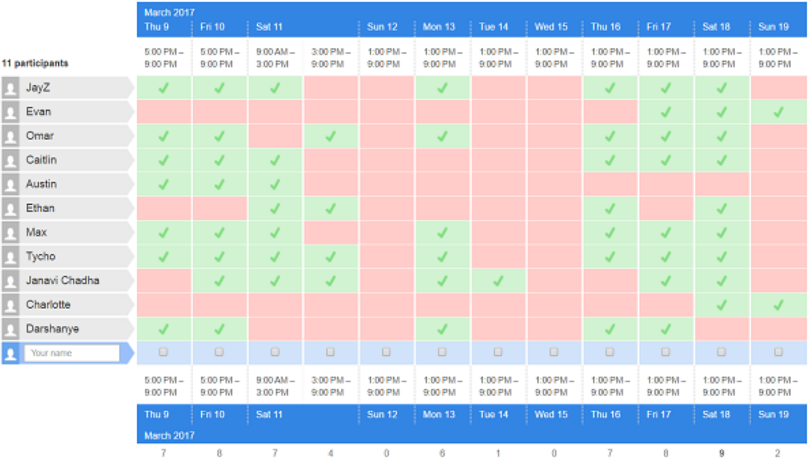
Extensions

- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

How would you schedule a meeting of 20 people?

How would you schedule a meeting of 20 people?



March 2017

	Thu 9	Fri 10	Sat 11		Sun 12	Mon 13	Tue 14	Wed 15	Thu 16	Fri 17	Sat 18	Sun 19
11 participants	5:00 PM – 9:00 PM	5:00 PM – 9:00 PM	9:00 AM – 3:00 PM	3:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM
JayZ	✓	✓	✓			✓			✓	✓	✓	
Evan										✓	✓	✓
Omar	✓	✓		✓		✓			✓	✓	✓	
Caitlin	✓	✓	✓						✓	✓	✓	
Austin	✓	✓	✓									
Ethan			✓	✓					✓		✓	
Max	✓	✓	✓			✓			✓	✓	✓	
Tycho	✓	✓	✓	✓		✓			✓	✓	✓	
Janavi Chadha		✓	✓	✓		✓	✓			✓	✓	
Charlotte											✓	✓
Darshanye	✓	✓				✓			✓	✓		
Your name	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	5:00 PM – 9:00 PM	5:00 PM – 9:00 PM	9:00 AM – 3:00 PM	3:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM	1:00 PM – 9:00 PM
	Thu 9	Fri 10	Sat 11		Sun 12	Mon 13	Tue 14	Wed 15	Thu 16	Fri 17	Sat 18	Sun 19
	7	8	7	4	0	6	1	0	7	8	**9**	2

March 2017

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

Other applications

Extensions

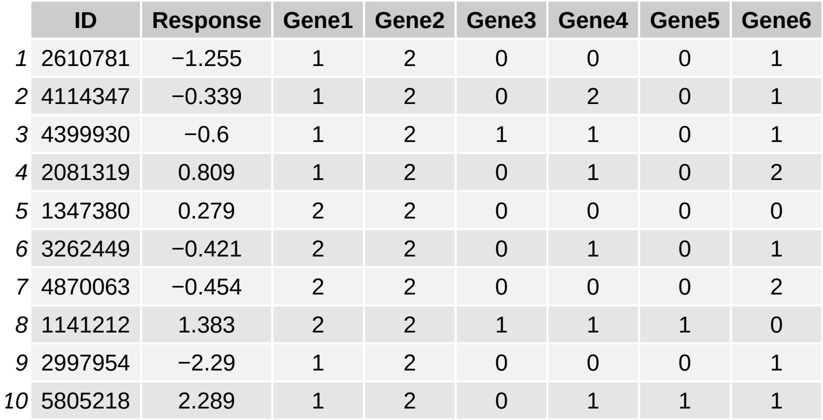
- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

UKBiobank

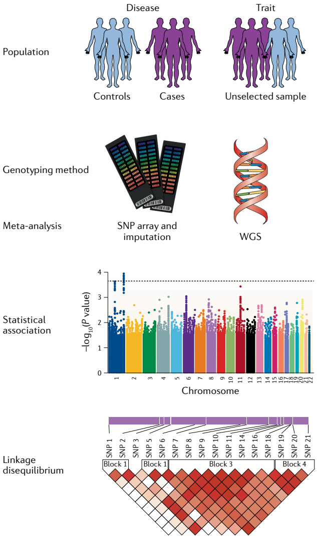
- Données de génotypage sont issues de 500 000 individus d’origine caucasienne recrutés au Royaume-Uni
- La puce UKBioBANK comporte plus de 800 000 SNPs
- Grand nombre de variables réponses (ex. maladie, densité minérale osseuse)
- Objectif: Quelles variables explicatives sont associées à la variable réponse?

Un échantillon



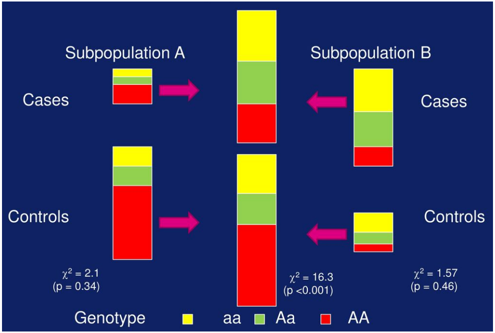
	ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	−1.255	1	2	0	0	0	1
2	4114347	−0.339	1	2	0	2	0	1
3	4399930	−0.6	1	2	1	1	0	1
4	2081319	0.809	1	2	0	1	0	2
5	1347380	0.279	2	2	0	0	0	0
6	3262449	−0.421	2	2	0	1	0	1
7	4870063	−0.454	2	2	0	0	0	2
8	1141212	1.383	2	2	1	1	1	0
9	2997954	−2.29	1	2	0	0	0	1
10	5805218	2.289	1	2	0	1	1	1

GWAS



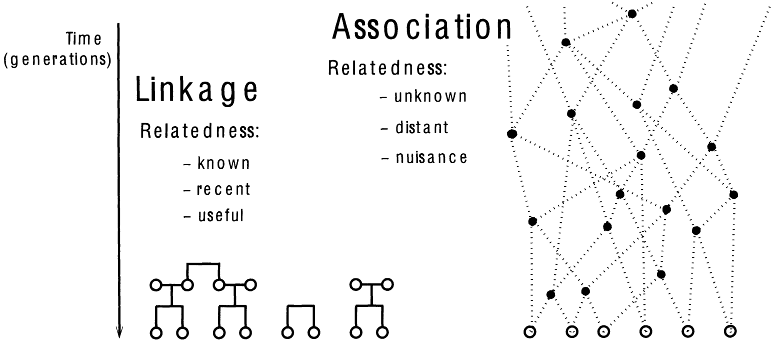
[1] Tam V. et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet (2019)

Confounding



Population structure

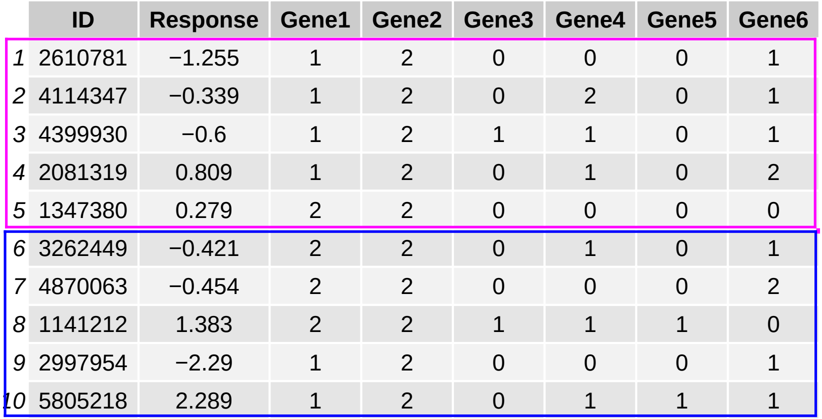
- Les GWAS comparent des individus non apparentés, mais «non apparentés» en fait signifie que les relations sont **inconnues** et présumées éloignées.



[1] Astle and Balding. Population structure and cryptic relatedness in genetic association studies. Statistical Science (2009)

Les observations ne sont pas indépendants

- Les observations sont **corrélées**, mais cette relation est souvent **inconnue**
- Cependant, elle peut être **estimé** à partir des données



	ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	−1.255	1	2	0	0	0	1
2	4114347	−0.339	1	2	0	2	0	1
3	4399930	−0.6	1	2	1	1	0	1
4	2081319	0.809	1	2	0	1	0	2
5	1347380	0.279	2	2	0	0	0	0
6	3262449	−0.421	2	2	0	1	0	1
7	4870063	−0.454	2	2	0	0	0	2
8	1141212	1.383	2	2	1	1	1	0
9	2997954	−2.29	1	2	0	0	0	1
10	5805218	2.289	1	2	0	1	1	1

La matrice de parenté (kinship)

- Soit *kinship* une liste de SNP utilisée pour estimer la matrice de parenté
- Soit $X_{kinship}$ une matrice de génotype normalisée $n \times q$.
- Une matrice de parenté ($\mathbf{\Phi}$) peut être calculée comme:

$$\mathbf{\Phi} = \frac{1}{q-1} X_{kinship} X_{kinship}^{\top} \tag{1}$$

Test d’association avec un modèle mixte linéaire (LMM)

$$\mathbf{Y} = \sum_{j=1}^{p} \beta_j \cdot \mathrm{SNP}_j + \mathbf{P} + \boldsymbol{\varepsilon} \tag{2}$$

$$\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\mathbf{I})$$

- σ^2 est la variance totale du phénotype
- $\eta \in [0, 1]$ est l’héritabilité du phénotype
- $\mathbf{Y} | (\eta, \sigma^2) \sim \mathcal{N}(\mathbf{0}, \eta\sigma^2\boldsymbol{\Phi} + (1-\eta)\sigma^2\mathbf{I})$

Régression ridge (Hoerl & Kennard 1970, Technometrics), Lasso (Tibshirani 1996, JRSSB)

- $\widehat{\boldsymbol{\beta}^{ridge}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_2^2$
- $\widehat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$

Lasso, ridge, ect. ne sont pas directement applicable au LMM

Procédure en deux étapes

- Étape 1: Ajuster un LMM sous l'hypothèse nul avec un seul effet aléatoire

$$\mathbf{Y} = \mathbf{P} + \boldsymbol{\varepsilon}$$

$$\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\mathbf{I})$$

Procédure en deux étapes

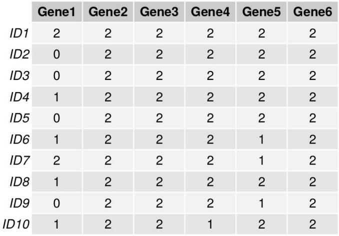
- Étape 1: Ajuster un LMM sous l'hypothèse nul avec un seul effet aléatoire

$$\mathbf{Y} = \mathbf{P} + \boldsymbol{\varepsilon}$$
$$\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\mathbf{I})$$

- Étape 2: Utilisez les résidus de l'étape 1 comme nouvelle réponse *indépendante*

Procédure en deux étapes

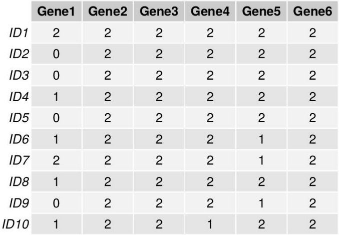
X_kinship



	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

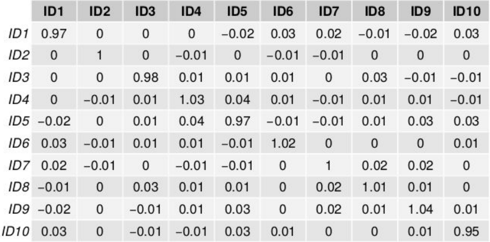
Procédure en deux étapes

X_kinship



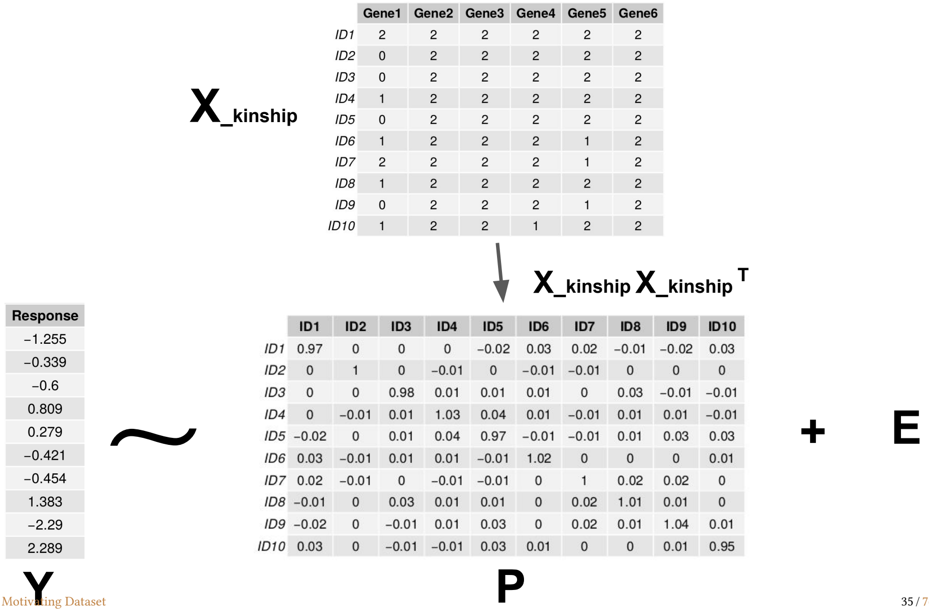
	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

X_kinship X_kinship ^{T}



	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

Procédure en deux étapes



X_kinship

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

$X\_kinship \; X\_kinship^T$

Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289

$\sim$

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

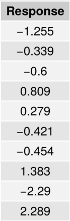
$+ \quad E$

Y

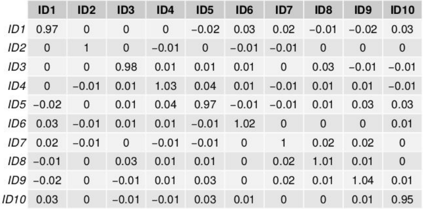
P

Procédure en deux étapes

Step 1: **Y** $\sim$ **P** $+ \; E_1$

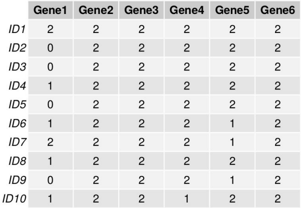


Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289



	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

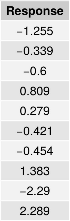
Step 2: Residuals from Step 1 $\sim$ $+ \; E_2$



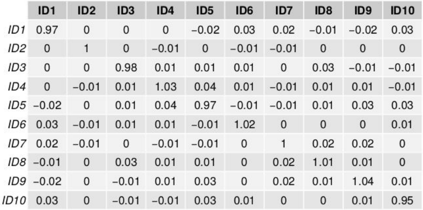
	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

Procédure en deux étapes

Step 1: **Y** ~ **P** + E_1

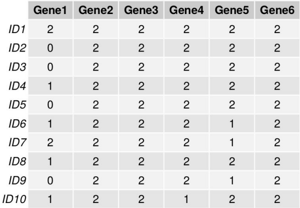


Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289



	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

Step 2: Residuals from Step 1 ~ + E_2



	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

- Dans les tests d'association, on sait qu'il souffre d'énormes pertes de puissance (Oualkacha et al. Gene. Epi. (2013))

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

Other applications

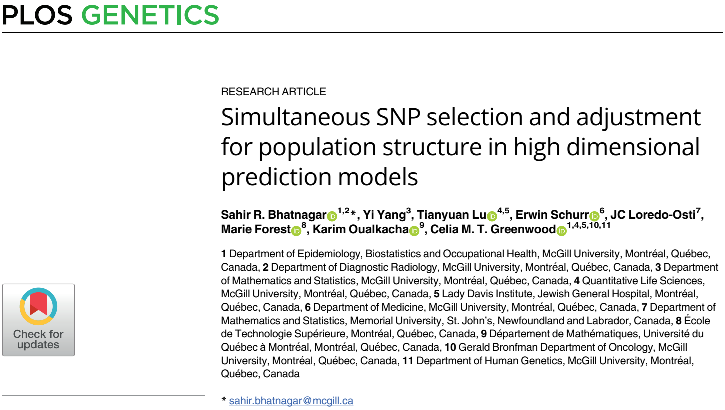
Extensions

- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

Notre proposition

- Nous proposons, `ggmix`, une procédure en **une seule étape** qui contrôle simultanément les populations structurées et effectue une sélection de variables dans les modèles mixtes linéaires



PLOS GENETICS

RESEARCH ARTICLE

Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models

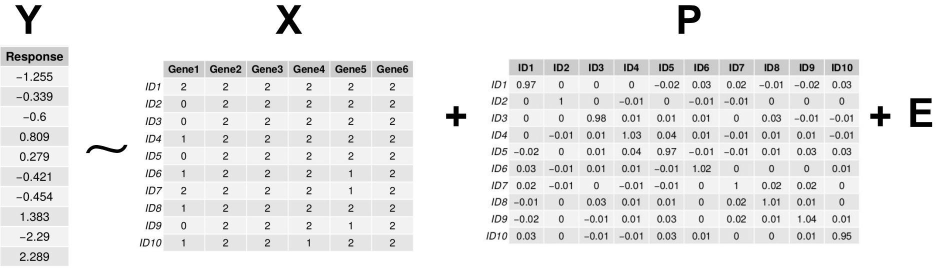
Sahir R. Bhatnagar[1,2]*, **Yi Yang**[3], **Tianyuan Lu**[4,5], **Erwin Schurr**[6], **JC Loredo-Osti**[7], **Marie Forest**[8], **Karim Oualkacha**[9], **Celia M. T. Greenwood**[1,4,5,10,11]

1 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada, **2** Department of Diagnostic Radiology, McGill University, Montréal, Québec, Canada, **3** Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada, **4** Quantitative Life Sciences, McGill University, Montréal, Québec, Canada, **5** Lady Davis Institute, Jewish General Hospital, Montréal, Québec, Canada, **6** Department of Medicine, McGill University, Montréal, Québec, Canada, **7** Department of Mathematics and Statistics, Memorial University, St. John's, Newfoundland and Labrador, Canada, **8** École de Technologie Supérieure, Montréal, Québec, Canada, **9** Département de Mathématiques, Université du Québec à Montréal, Montréal, Québec, Canada, **10** Gerald Bronfman Department of Oncology, McGill University, Montréal, Québec, Canada, **11** Department of Human Genetics, McGill University, Montréal, Québec, Canada

* sahir.bhatnagar@mcgill.ca

[1] R package: `sahirbhatnagar.com/ggmix`, `https://cran.r-project.org/package=ggmix`

ggmix: une procédure en **une seule étape**



Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

[1] R package: sahirbhatnagar.com/ggmix, https://cran.r-project.org/package=ggmix

Data and Model

- Phenotype: $\mathbf{Y} = (y_1, \dots, y_n) \in \mathbb{R}^n$
- SNPs: $\mathbf{X} = (\mathbf{X}_1; \dots, \mathbf{X}_n)^T \in \mathbb{R}^{n \times p}$, where $p \gg n$
- Twice the Kinship matrix or Realized Relationship matrix: $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$
- Regression Coefficients: $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$
- Polygenic random effect: $\mathbf{P} = (P_1, \dots, P_n) \in \mathbb{R}^n$
- Error: $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$
- We consider the following LMM with a single random effect:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{P} + \boldsymbol{\varepsilon}$$

$$\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\mathbf{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\mathbf{I})$$

- σ^2 is the phenotype total variance
- $\eta \in [0, 1]$ is the phenotype heritability (narrow sens)
- $\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \eta\sigma^2\mathbf{\Phi} + (1-\eta)\sigma^2\mathbf{I})$

Likelihood

- The negative log-likelihood is given by

$$-\ell(\boldsymbol{\Theta}) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(\mathbf{V})\right) + \frac{1}{2\sigma^2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)^T \mathbf{V}^{-1}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)$$

$$\mathbf{V} = \eta\boldsymbol{\Phi} + (1-\eta)\mathbf{I}$$

- Assume the spectral decomposition of $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

- $\mathbf{U}$ is an $n \times n$ orthogonal matrix and $\mathbf{D}$ is an $n \times n$ diagonal matrix
- One can write

$$\mathbf{V} = \mathbf{U}(\eta\mathbf{D} + (1-\eta)\mathbf{I})\mathbf{U}^\top = \mathbf{U}\mathbf{W}\mathbf{U}^\top$$

  with $\mathbf{W} = \operatorname{diag}\left(w_i\right)_{i=1}^{n}$, $w_i = \eta\mathbf{D}_{ii} + (1-\eta)$

Likelihood

- Projection of $\mathbf{Y}$ (and columns of $\mathbf{X}$) into Span($\mathbf{U}$) leads to a simplified correlation structure for the transformed data: $\tilde{\mathbf{Y}} = \mathbf{U}^\top \mathbf{Y}$
- $\tilde{\mathbf{Y}} | (\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 \mathbf{W})$, with $\tilde{\mathbf{X}} = \mathbf{U}^\top \mathbf{X}$
- The negative log-likelihood can then be expressed as

$$-\ell(\boldsymbol{\Theta}) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \log\left(w_i\right) + \frac{1}{2\sigma^2} \left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)^T \mathbf{W}^{-1} \left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)$$

Likelihood

- Projection of $\mathbf{Y}$ (and columns of $\mathbf{X}$) into Span($\mathbf{U}$) leads to a simplified correlation structure for the transformed data: $\tilde{\mathbf{Y}} = \mathbf{U}^\top \mathbf{Y}$
- $\tilde{\mathbf{Y}} | (\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 \mathbf{W})$, with $\tilde{\mathbf{X}} = \mathbf{U}^\top \mathbf{X}$
- The negative log-likelihood can then be expressed as

$$-\ell(\boldsymbol{\Theta}) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \log\left(w_i\right) + \frac{1}{2\sigma^2} \left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)^T \mathbf{W}^{-1} \left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)$$

- For fixed σ^2 and η, solving for $\boldsymbol{\beta}$ is a weighted least squares problem

Penalized Maximum Likelihood Estimator

- Define the objective function:

$$Q_\lambda(\boldsymbol{\Theta}) = -\ell(\boldsymbol{\Theta}) + \lambda \sum_j p_j(\beta_j)$$

- $p_j(\cdot)$ is a penalty term on $\beta_1, \ldots, \beta_p$
- An estimate of the model parameters $\widehat{\boldsymbol{\Theta}}_\lambda$ is obtained by

$$\widehat{\boldsymbol{\Theta}}_\lambda = \arg\min_{\boldsymbol{\Theta}} Q_\lambda(\boldsymbol{\Theta})$$

Block Relaxation (De Leeuw, 1994)

To solve for the optimization problem we use a block relaxation technique

```
Set $k \leftarrow 0$, initial values for the parameter vector $\boldsymbol{\Theta}^{(0)}$ and $\epsilon$;
for $\lambda \in \{\lambda_{max}, \ldots, \lambda_{min}\}$ do
    repeat
        $For\ j = 1, \ldots, p,\ \beta_j^{(k+1)} \leftarrow \arg\min_{\beta_j} Q_\lambda \left(\boldsymbol{\beta}_{-j}^{(k)}, \eta^{(k)}, \sigma^{2\ (k)}\right)$
        $\eta^{(k+1)} \leftarrow \arg\min_{\eta} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta, \sigma^{2\ (k)}\right)$
        $\sigma^{2\ (k+1)} \leftarrow \arg\min_{\sigma^2} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta^{(k+1)}, \sigma^2\right)$
        $k \leftarrow k + 1$
    until convergence criterion is satisfied: $||\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Theta}^{(k)}||_2 < \epsilon$;
end
```

Algorithm 1: Block Relaxation Algorithm

Coordinate Gradient Descent Method

- We take advantage of smoothness of $\ell(\boldsymbol{\Theta})$
- We approximate $Q_\lambda(\boldsymbol{\Theta})$ by a strictly convex quadratic function (using gradient)
- We use CGD to calculate a descent direction
- To achieve the descent property for the objective function, we employ further line search

[1] Tseng P& Yun S. Math. Program., Ser. B, (2009)

Coordinate Gradient Descent Method

- We take advantage of smoothness of $\ell(\mathbf{\Theta})$
- We approximate $Q_\lambda(\mathbf{\Theta})$ by a strictly convex quadratic function (using gradient)
- We use CGD to calculate a descent direction
- To achieve the descent property for the objective function, we employ further line search

Theorem [Convergence] [1]**:**
If $\{\mathbf{\Theta}^{(k)}, k = 0, 1, 2, \ldots\}$ is a sequence of iterates generated by the iteration map of Algorithm 1, then each cluster point (i.e. limit point) of $\{\mathbf{\Theta}^{(k)}, k = 0, 1, 2, \ldots\}$ is a stationary point of $Q_\lambda(\mathbf{\Theta})$

[1] Tseng P& Yun S. Math. Program., Ser. B, (2009)

Choice of the tuning parameter

- We use the BIC:

$$BIC_\lambda = -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \widehat{\eta}) + c \cdot \widehat{df}_\lambda$$

- $\widehat{df}_\lambda$ is the number of non-zero elements in $\widehat{\boldsymbol{\beta}}_\lambda$ plus two [1]
- Several authors [2] have used this criterion for variable selection in mixed models with $c = \log n$
- Other authors [3] have proposed $c = \log(\log(n)) * \log(n)$

[1] Zou et al. The Annals of Statistics, (2007)

[2] Bondell et al. Biometrics (2010)

[3] Wang et al. JRSS(Ser. B), (2009)

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

Other applications

Extensions

- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

Simulation study

- We simulated data from the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{P} + \boldsymbol{\varepsilon}$
- We used heritability $\eta = \{0.1, 0.3\}$, number of covariates $p = 5,000$, number of *kinship* SNPs $k = 10,000$, percentage of *causal* SNPs $c = \{0\%, 1\%\}$ and $\sigma^2 = 1$.
- In addition to these parameters, we also varied the amount of overlap between the *causal* list and the *kinship* list:
  1. None of the *causal* SNPs are included in *kinship* set.
  2. All of the *causal* SNPs are included in the *kinship* set.
- These were meant to contrast the model behavior when causal SNPs are included in both the main effects and random effects vs. when the causal SNPs are only included in the main effects.
- These scenarios are motivated by the current standard of practice in GWAS where the candidate marker is excluded from the calculation of the kinship matrix.
- This approach becomes much more difficult to apply in large-scale multivariable models where there is likely to be overlap between the variables in the design matrix and kinship matrix.

Simulation study results

- Both the `lasso+PC` and `twostep` selected more false positives compared to `ggmix`
- Overall, we observed that variable selection results and RMSE for `ggmix` were similar regardless of whether the causal SNPs were in the kinship matrix or not.
- This result is encouraging since in practice the kinship matrix is constructed from a random sample of SNPs across the genome, some of which are likely to be causal, particularly in polygenic traits.
- In particular, our simulation results show that the principal component adjustment method may not be the best approach to control for confounding by population structure, particularly when variable selection is of interest.

Real data applications

1. **UK Biobank**
   - 10,000 LD-pruned SNPs (Essentially un-correlated variables) to predict standing height in 18k related individuals
   - Standing height is highly polygenic (many variables associated with response)

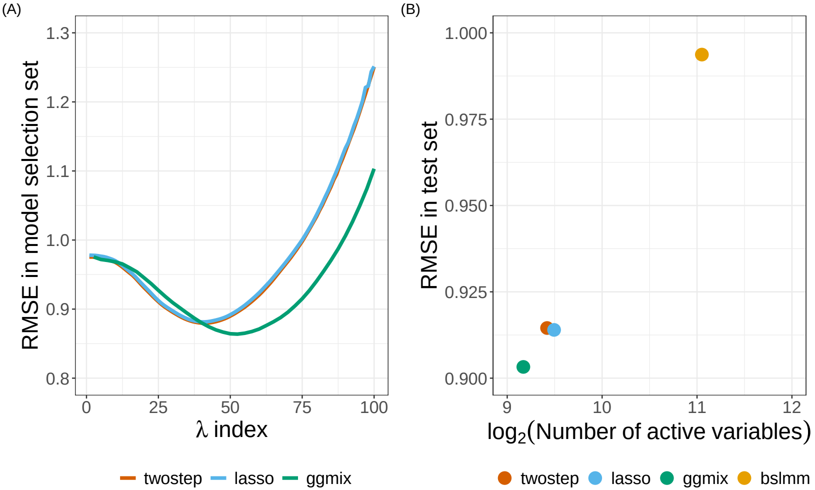
Real data applications

1. **UK Biobank**
   - 10,000 LD-pruned SNPs (Essentially un-correlated variables) to predict standing height in 18k related individuals
   - Standing height is highly polygenic (many variables associated with response)
2. **GAW20 Simulated dataset**
   - 50,000 SNPs (all on chromosome 1) to predict high-density lipoproteins in 679 related individuals
   - Not much correlation between causal SNP and others
   - Very sparse signals (only 1 causal variant)

Real data applications

1. **UK Biobank**
   - 10,000 LD-pruned SNPs (Essentially un-correlated variables) to predict standing height in 18k related individuals
   - Standing height is highly polygenic (many variables associated with response)
2. **GAW20 Simulated dataset**
   - 50,000 SNPs (all on chromosome 1) to predict high-density lipoproteins in 679 related individuals
   - Not much correlation between causal SNP and others
   - Very sparse signals (only 1 causal variant)
3. **Mouse Crosses**
   - Find loci associated with mouse sensitivity to mycobacterial infection
   - 189 samples, and 625 microsatellite markers
   - Highly correlated variables

Results: UK Biobank

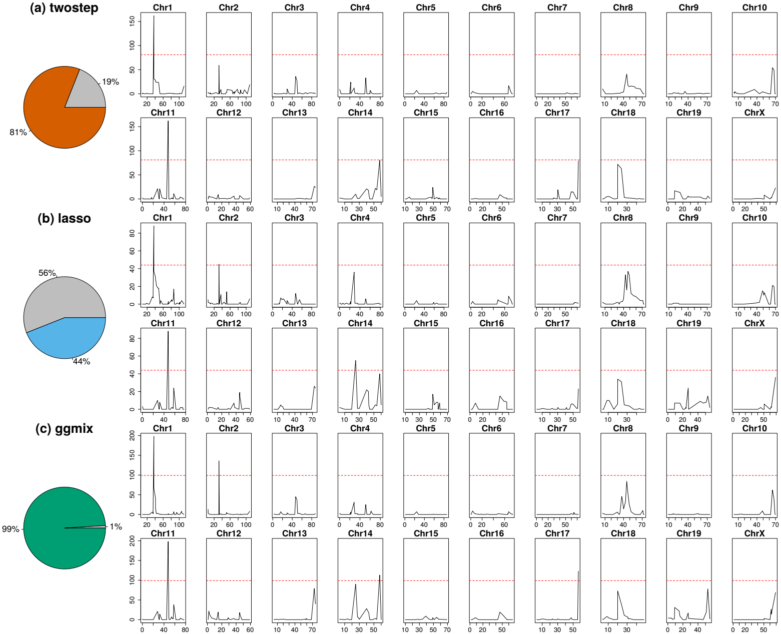


Results: GAW20

Method	Median number of active variables (Inter-quartile range)	RMSE (SD)
twostep	1 (1 - 11)	0.3604 (0.0242)
lasso	1 (1 - 15)	0.3105 (0.0199)
ggmix	1 (1 - 12)	0.3146 (0.0210)
BSLMM	40,737 (39,901 - 41,539)	0.2503 (0.0099)

Table: Summary of model performance based on 200 GAW20 simulations. Five-fold cross-validation root-mean-square error was reported for each simulation replicate.

Results: Mouse crosses

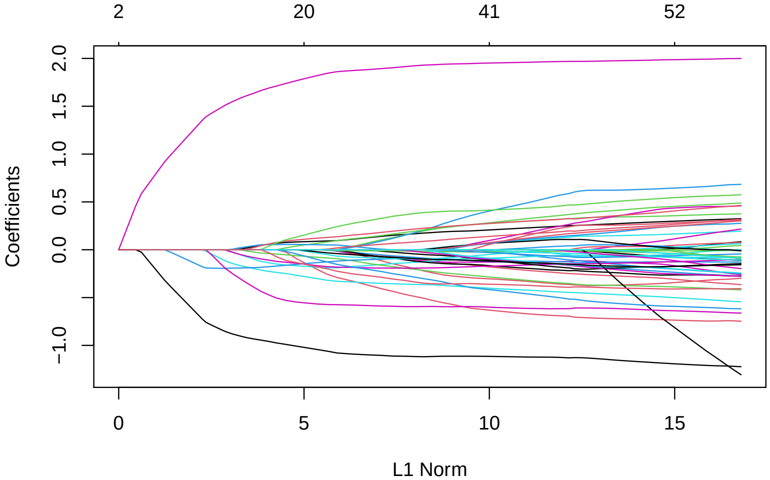


Discussion

- La procédure en deux étapes conduit à un grand nombre de faux positifs et de faux négatifs
- L'ajustement de la composante principale dans `lasso` peut ne pas être suffisant pour contrôler la confusion, en particulier lorsqu'il y a beaucoup de corrélation entre les observations
- `ggmix` fonctionne bien même lorsque les variables causales sont utilisées dans le calcul de la matrice de parenté
- `ggmix` a montré la plus grande amélioration par rapport à `twostep` et `lasso` quand il y avait des variables hautement corrélées avec beaucoup de structure (exemple de croix de souris)

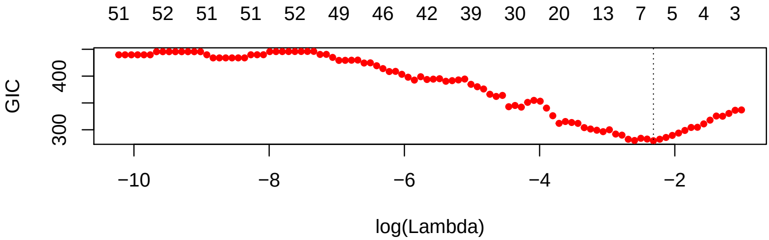
ggmix R package

```r
library(ggmix)
data("admixed")
fit <- ggmix(x = admixed$xtrain,
             y = admixed$ytrain,
             kinship = admixed$kin_train)
plot(fit)
```



ggmix R package

```r
hdbic <- gic(fit)
plot(hdbic)
```



```r
coef(hdbic, type = "nonzero")
```

```
##                       1
## (Intercept) -0.03598164
## X302        -0.17617815
## X524         1.34917874
## X538        -0.72073279
## eta          0.99000000
## sigma2       1.60477653
```

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

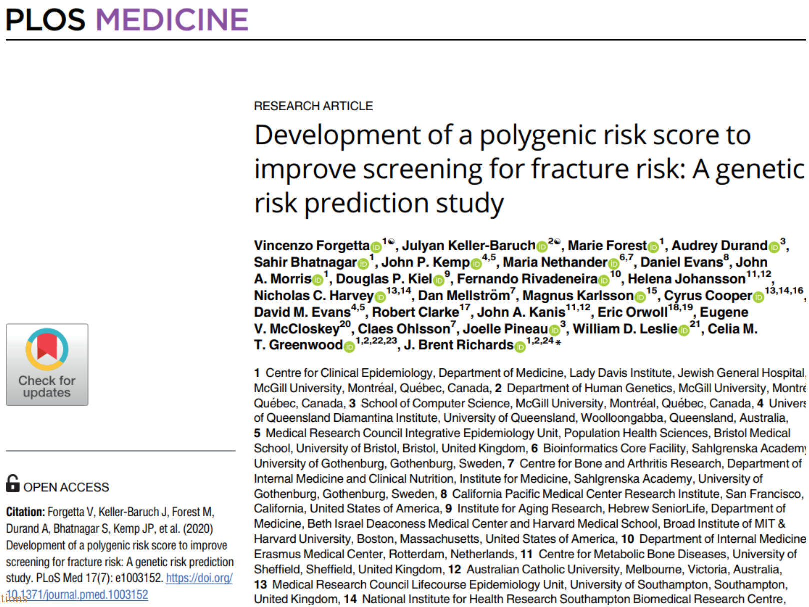
Other applications

Extensions

- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

gSOS



PLOS MEDICINE

RESEARCH ARTICLE

Development of a polygenic risk score to improve screening for fracture risk: A genetic risk prediction study

Vincenzo Forgetta[1☯], Julyan Keller-Baruch[2☯], Marie Forest[1], Audrey Durand[3], Sahir Bhatnagar[1], John P. Kemp[4,5], Maria Nethander[6,7], Daniel Evans[8], John A. Morris[1], Douglas P. Kiel[9], Fernando Rivadeneira[10], Helena Johansson[11,12], Nicholas C. Harvey[13,14], Dan Mellström[7], Magnus Karlsson[15], Cyrus Cooper[13,14,16], David M. Evans[4,5], Robert Clarke[17], John A. Kanis[11,12], Eric Orwoll[18,19], Eugene V. McCloskey[20], Claes Ohlsson[7], Joelle Pineau[3], William D. Leslie[21], Celia M. T. Greenwood[1,2,22,23], J. Brent Richards[1,2,24]*

1 Centre for Clinical Epidemiology, Department of Medicine, Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada, **2** Department of Human Genetics, McGill University, Montré Québec, Canada, **3** School of Computer Science, McGill University, Montréal, Québec, Canada, **4** Univers of Queensland Diamantina Institute, University of Queensland, Woolloongabba, Queensland, Australia, **5** Medical Research Council Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom, **6** Bioinformatics Core Facility, Sahlgrenska Academy University of Gothenburg, Gothenburg, Sweden, **7** Centre for Bone and Arthritis Research, Department of Internal Medicine and Clinical Nutrition, Institute for Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, **8** California Pacific Medical Center Research Institute, San Francisco, California, United States of America, **9** Institute for Aging Research, Hebrew SeniorLife, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Broad Institute of MIT & Harvard University, Boston, Massachusetts, United States of America, **10** Department of Internal Medicine Erasmus Medical Center, Rotterdam, Netherlands, **11** Centre for Metabolic Bone Diseases, University of Sheffield, Sheffield, United Kingdom, **12** Australian Catholic University, Melbourne, Victoria, Australia, **13** Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, Southampton, United Kingdom, **14** National Institute for Health Research Southampton Biomedical Research Centre,

Check for updates

OPEN ACCESS

Citation: Forgetta V, Keller-Baruch J, Forest M, Durand A, Bhatnagar S, Kemp JP, et al. (2020) Development of a polygenic risk score to improve screening for fracture risk: A genetic risk prediction study. PLoS Med 17(7): e1003152. https://doi.org/10.1371/journal.pmed.1003152

gSOS

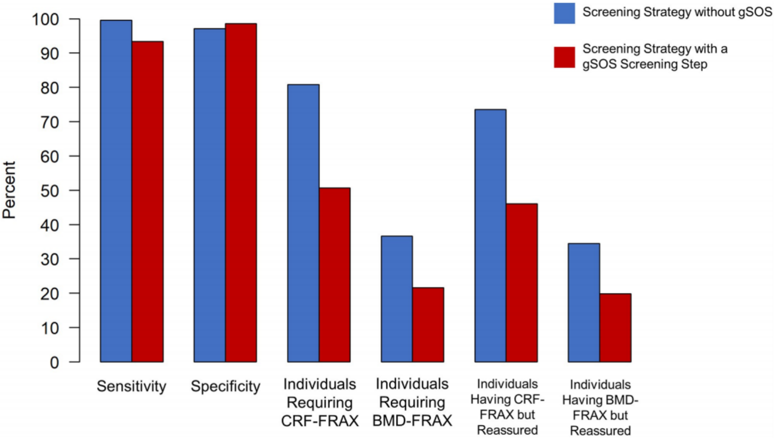


Fig 6. Performance characteristics of screening with and without a gSOS screening step. BMD-FRAX, bone-mineral-density-based Fracture Risk Assessment Tool; CRF-FRAX, clinical-risk-factor-based Fracture Risk Assessment Tool.

https://doi.org/10.1371/journal.pmed.1003152.g006

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

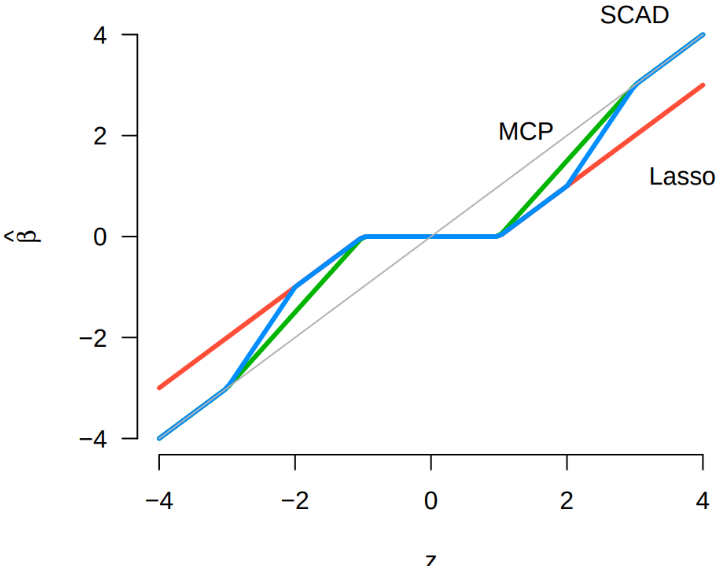
Other applications

Extensions

- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

SCAD (Fan et Li, JASA, 2001), MCP (Zhang, Ann. Stat., 2010)



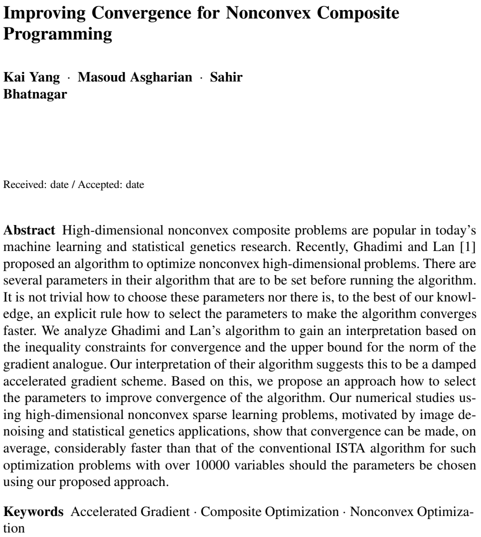
Computational challenges

- Past approaches for optimization for SCAD/MCP relies upon descent method, first- or second- order

- e.g., `sparsenet` (Mazumder et al. 2011) uses coordinate descent with full step size, whose coordinate update cycles through $\tilde{\beta}_j = S_{\gamma_k}\left(\sum_{i=1}^{n}\left(y_i - \tilde{y}_i^j\right)x_{ij}, \lambda_\ell\right)$, where $\tilde{y}_i^j = \sum_{k\neq j} x_{ik}\tilde{\beta}_k$

- However, coordinate descent is difficult to vectorize, and rate of convergence is difficult of establish – though past literature suggests $O(1/k)$ rate of convergence for ISTA

Our proposal: Accelerated gradient (AG) method



Improving Convergence for Nonconvex Composite Programming

Kai Yang · Masoud Asgharian · Sahir Bhatnagar

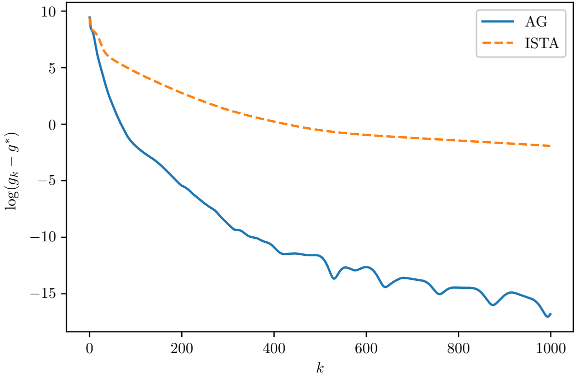
Received: date / Accepted: date

Abstract High-dimensional nonconvex composite problems are popular in today's machine learning and statistical genetics research. Recently, Ghadimi and Lan [1] proposed an algorithm to optimize nonconvex high-dimensional problems. There are several parameters in their algorithm that are to be set before running the algorithm. It is not trivial how to choose these parameters nor there is, to the best of our knowledge, an explicit rule how to select the parameters to make the algorithm converges faster. We analyze Ghadimi and Lan's algorithm to gain an interpretation based on the inequality constraints for convergence and the upper bound for the norm of the gradient analogue. Our interpretation of their algorithm suggests this to be a damped accelerated gradient scheme. Based on this, we propose an approach how to select the parameters to improve convergence of the algorithm. Our numerical studies using high-dimensional nonconvex sparse learning problems, motivated by image denoising and statistical genetics applications, show that convergence can be made, on average, considerably faster than that of the conventional ISTA algorithm for such optimization problems with over 10000 variables should the parameters be chosen using our proposed approach.

Keywords Accelerated Gradient · Composite Optimization · Nonconvex Optimization

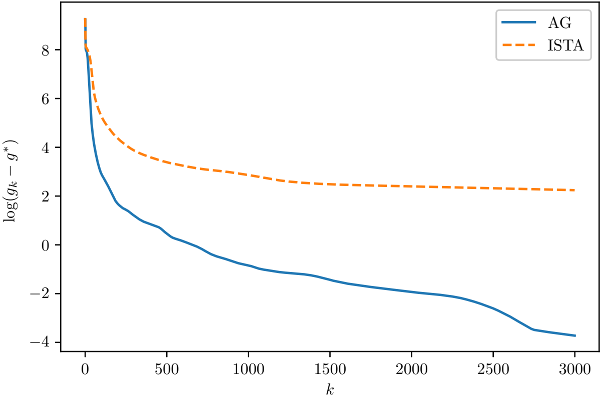
[1] https://arxiv.org/abs/2009.10629

Numerical Study for SCAD



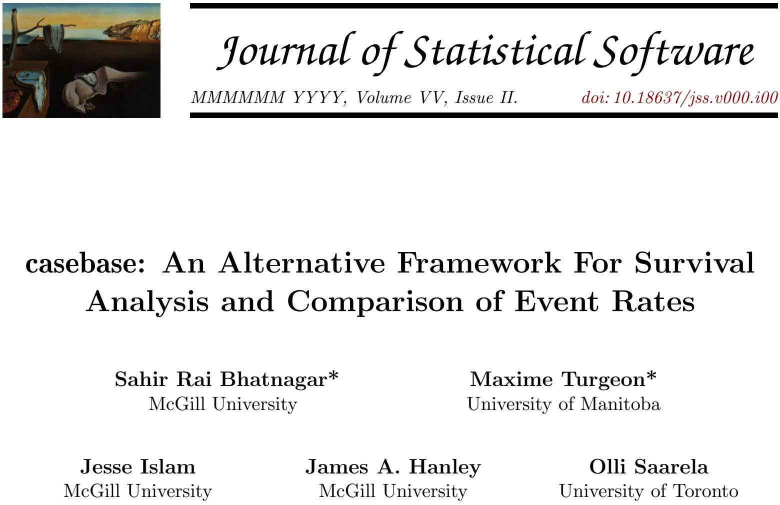
$\mathbf{x}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{I})$, $\varepsilon_i \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$, $\mathbf{y} = \mathbf{X}\boldsymbol{\tau}_{\text{generate}} + \boldsymbol{\varepsilon}$, $\sigma^2 = \frac{\left\|\boldsymbol{\tau}_{\text{generate}}\right\|^2}{3}$, $\boldsymbol{\tau}_{\text{generate}} \in \mathbb{R}^{10006}$ is a sparse constant vector with 6 values of 1.23(intercept)$, 3, 4, 5, 6, 59$ as true effect coefficients and 10000 values of 0. Start point: $\boldsymbol{\tau}_0 = \mathbf{1}_{10006}$, $a = 3.7,\ \lambda = 0.6$.

Numerical Study for MCP



Simulation settings here is same as before in SCAD, $\gamma = 2.5,\ \lambda = 0.6$.

casebase



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

casebase: An Alternative Framework For Survival Analysis and Comparison of Event Rates

Sahir Rai Bhatnagar*
McGill University

Maxime Turgeon*
University of Manitoba

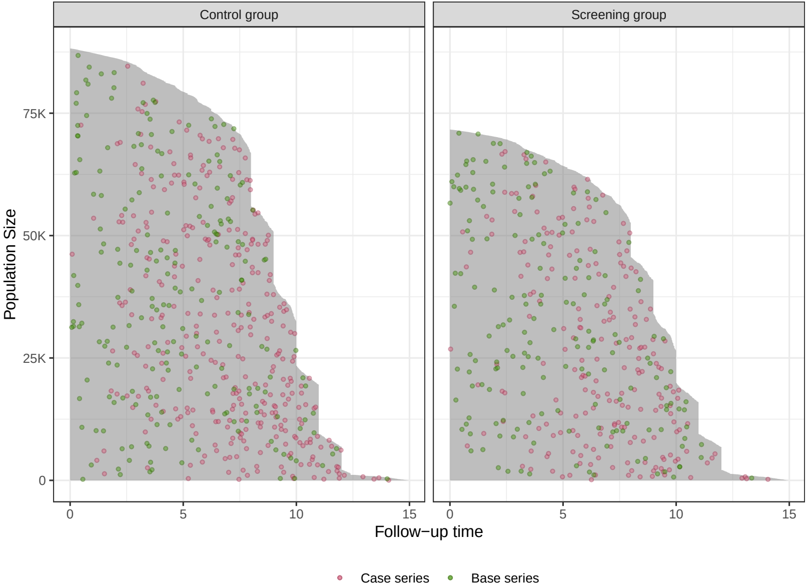
Jesse Islam
McGill University

James A. Hanley
McGill University

Olli Saarela
University of Toronto

[1] https://arxiv.org/abs/2009.10264, https://cran.r-project.org/package=casebase

Case-base sampling



Case-base sampling

- The unit of analysis is a person-moment.
- Case-base sampling reduces the model fitting to a familiar logistic regression.
- The sampling process is taken into account using an offset term.
- By sampling a large base series, the information loss eventually becomes negligible.
- This framework can easily be used with time-varying covariates (e.g. time-varying exposure). We can fit any hazard λ of the following form:

$$\log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X$$

- Different choices of the function g leads to familiar parametric families:
  - Exponential: g is constant.
  - Gompertz: $g(t; \alpha) = \alpha t$.
  - Weibull: $g(t; \alpha) = \alpha \log t$

Classical Methods

Betting on Sparsity

A Thought Experiment

Motivating Dataset

Notre proposition

Résultats

Other applications

Extensions

- Fonctions de pénalité non convexes
- Analyse de survie

Orientations futures

Orientations futures

- ggmix est limité par le nombre d'individus (ne s'applique pas à l'ensemble de la cohorte UK Biobank de 500k) → approximations de rang inférieur de la matrice de parenté
- Problèmes de mémoire lorsque le nombre de covariables dans le modèle dépasse 50k → stratégies de mappage de mémoire (par exemple biglasso de Zeng et Breheny (2017))
- Extension aux données multivariées, longitudinales, combinaisons de plusieurs cohortes → Plusieurs effets aléatoires.

Remerciements

CRSNG RGPIN-2020-05133

- Kai Yang: Non-convex optimization
- Jesse Islam: High-dimensional survival analysis

Kai Yang, PhD (c)

Jesse Islam, PhD (c)

NSERC
CRSNG

Remerciements

MiCM

- Julien St-Pierre: LMM with multiple random effects, longitudinal data, combining multiple cohorts

Julien St-Pierre, PhD (c)

MiCM McGill initiative in Computational Medicine

Remerciements

CIHR Project Grant, CANSSI CRT

- Zeyu Bian: Low-rank approximations, memory mapping
- Mohan Zhao: Multivariate outcomes and matrix covariates

Zeyu Bian, PhD (c)

Mohan Zhao, BSc (c)

CIHR IRSC

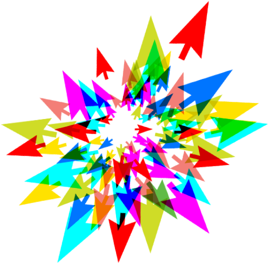
CANSSI INCASS

Remerciements

- Masoud Asgharian (McGill)
- Tianyuan Lu (McGill)
- Yi Yang (McGill)
- Karim Oualkacha (UQÀM)
- Celia Greenwood (Lady Davis Institute)
- Erica Moodie (McGill)
- James Hanley (McGill)
- Maxime Turgeon (UManitoba)
- Olli Saarela (UofT)
- Luda Diatchenko (McGill)
- UK Biobank Resource under project number 27449. We appreciate the generosity of UK Biobank volunteers

Calcul Québec

compute canada | calcul canada



References

1. **Yang K**, Asgharian M, Bhatnagar SR (2020+). Improving Rate of Convergence for Nonconvex Composite Programming. *Submitted to Optimization Letters*. https://arxiv.org/abs/2009.10629.
2. Bhatnagar SR, Turgeon M, **Islam J**, Hanley JA, Saarela O (2020+). casebase: An Alternative Framework For Survival Analysis and Comparison of Event Rates. *Submitted to Journal of Statistical Software*. https://arxiv.org/abs/2009.10264.
3. Bhatnagar SR, Yang Y, Lu T, Schurr E, Loredo-Osti JC, Forest M, Oualkacha K, Greenwood CMT (2020). Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. *PLoS Genetics* 16(5): e1008766. DOI 10.1371/journal.pgen.1008766.

sahirbhatnagar.com

Session Info

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 20.10

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.10.so

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ggmix_0.0.1 knitr_1.31

loaded via a namespace (and not attached):
 [1] lattice_0.20-41  codetools_0.2-16 glmnet_4.1-1     foreach_1.5.1
 [5] grid_4.0.2       magrittr_2.0.1   evaluate_0.14    highr_0.8
 [9] stringi_1.5.3    Matrix_1.2-18    splines_4.0.2    iterators_1.0.13
[13] tools_4.0.2      stringr_1.4.0    survival_3.2-3   xfun_0.21
[17] compiler_4.0.2   shape_1.4.5
```