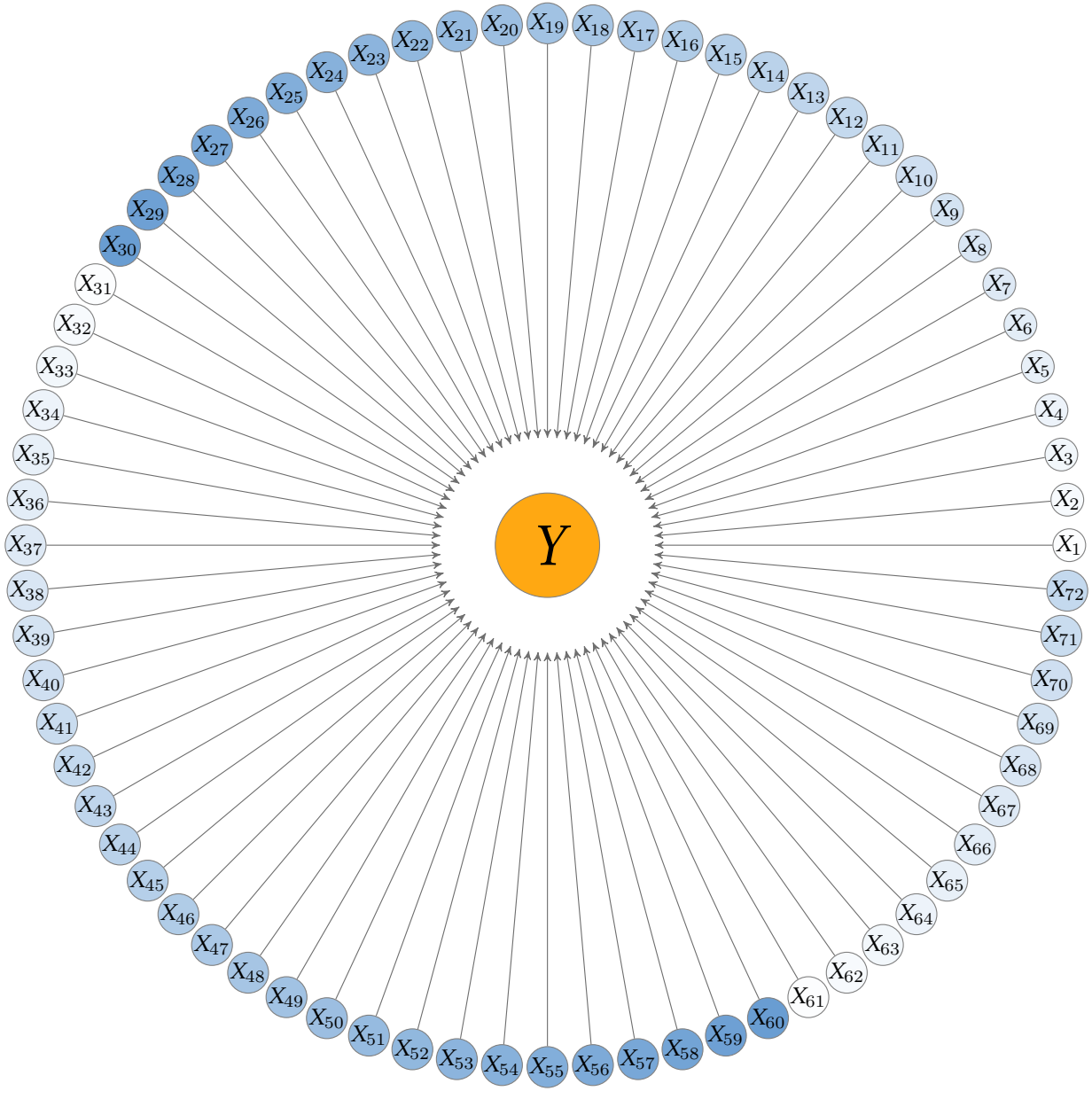
Avances dans les modèles mixtes pour la prévision et la sélection des variables dans les données de grande dimension

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

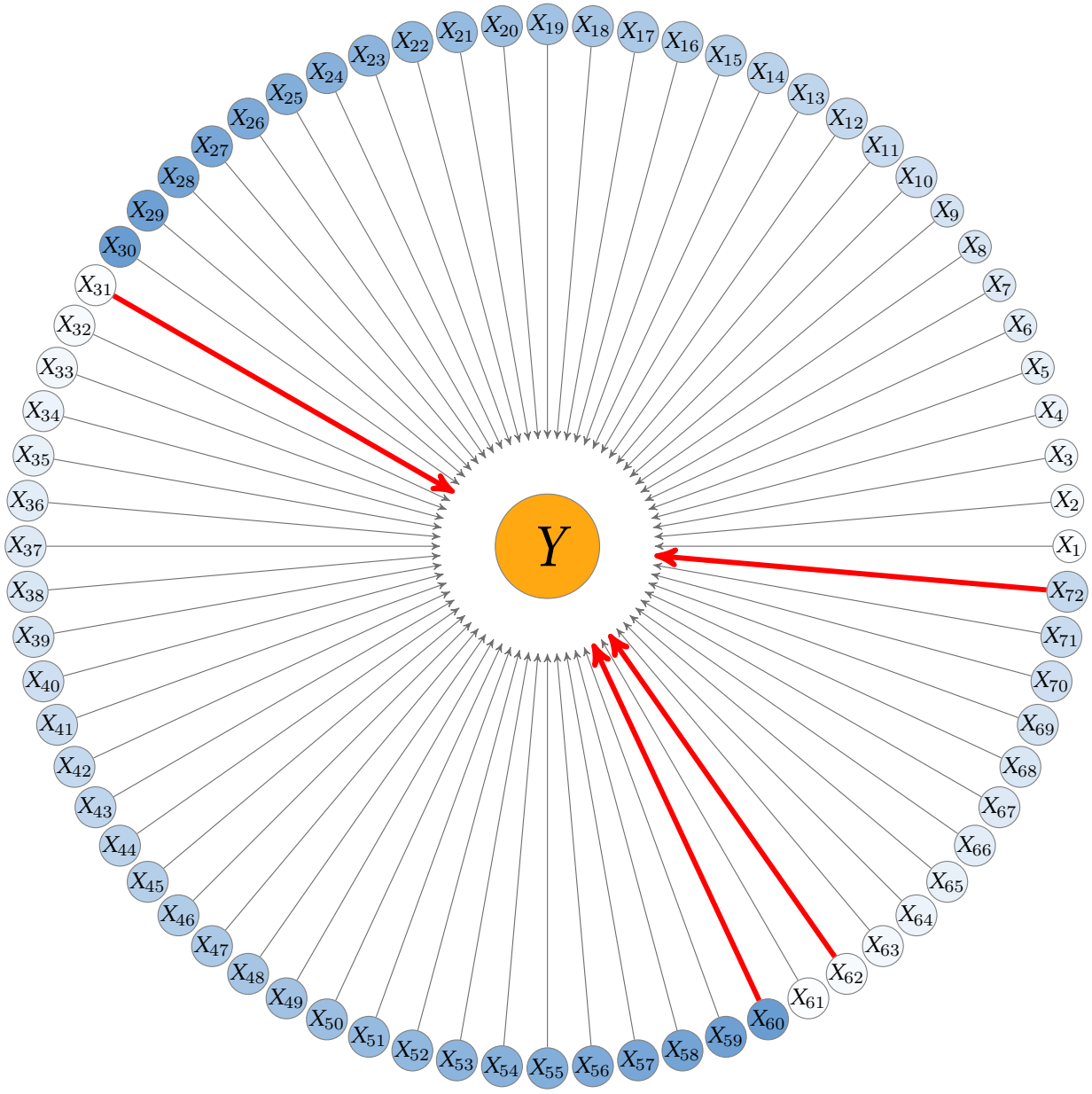
sahirbhatnagar.com

24 novembre 2020

Miser sur la sparsité



Miser sur la sparsité



Miser sur la sparsité

Utilisez une procédure qui fonctionne bien pour les problèmes sparse, car aucune procédure ne fonctionne bien pour les problèmes denses.[1]

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

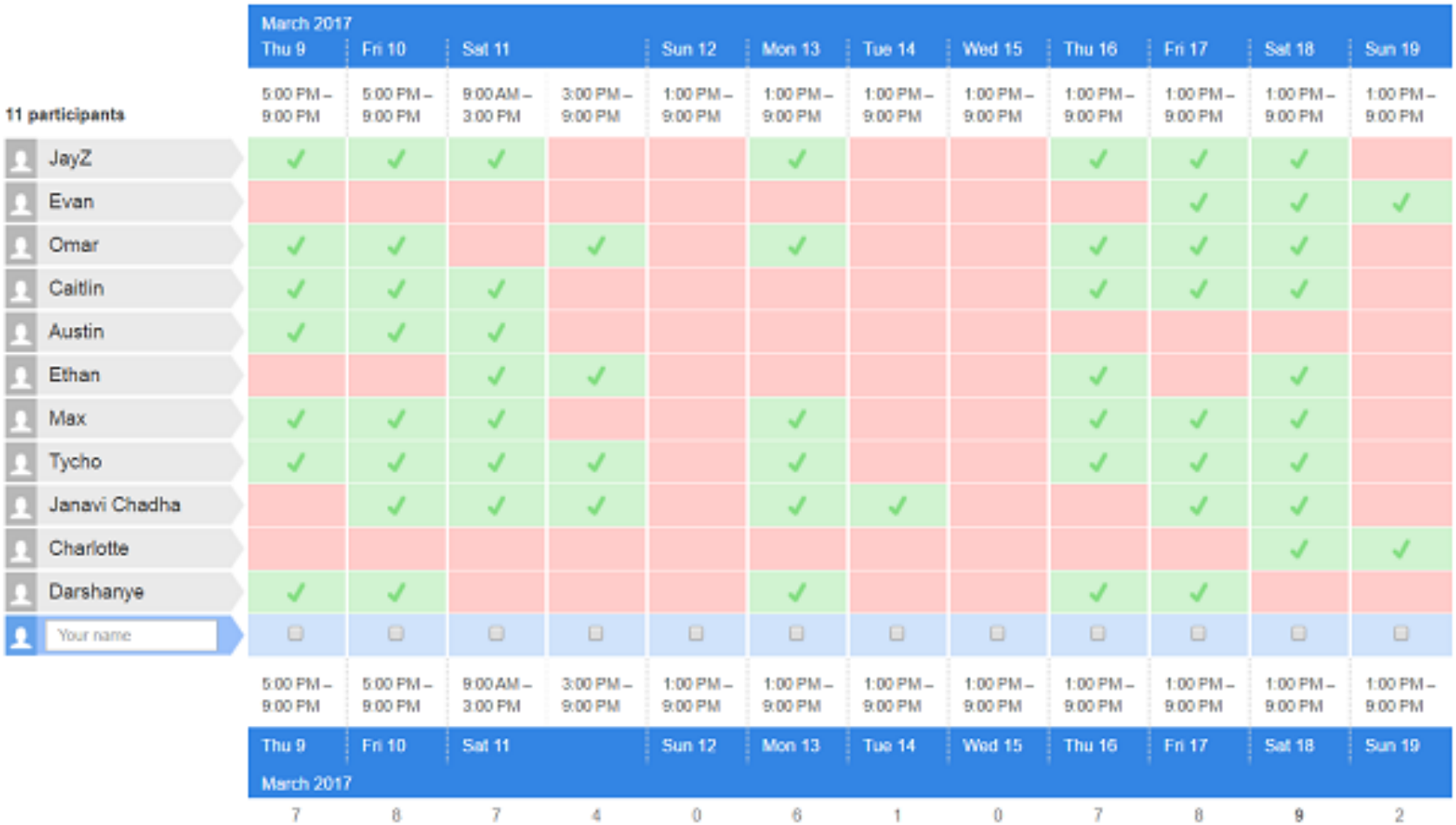
Miser sur la sparsité

Utilisez une procédure qui fonctionne bien pour les problèmes sparse, car aucune procédure ne fonctionne bien pour les problèmes denses.[1]

- Un modèle statistique sparse est un modèle pour lequel seulement un petit nombre de variables explicatives jouent un rôle important.
- Hypothèse de parcimonie: peu de variables sont pertinentes pour les données de grande dimension ($N << p$).
- β est "creux"
- Les modèles sparse peuvent être plus rapides à calculer, plus faciles à comprendre et produire des prédictions plus stables.

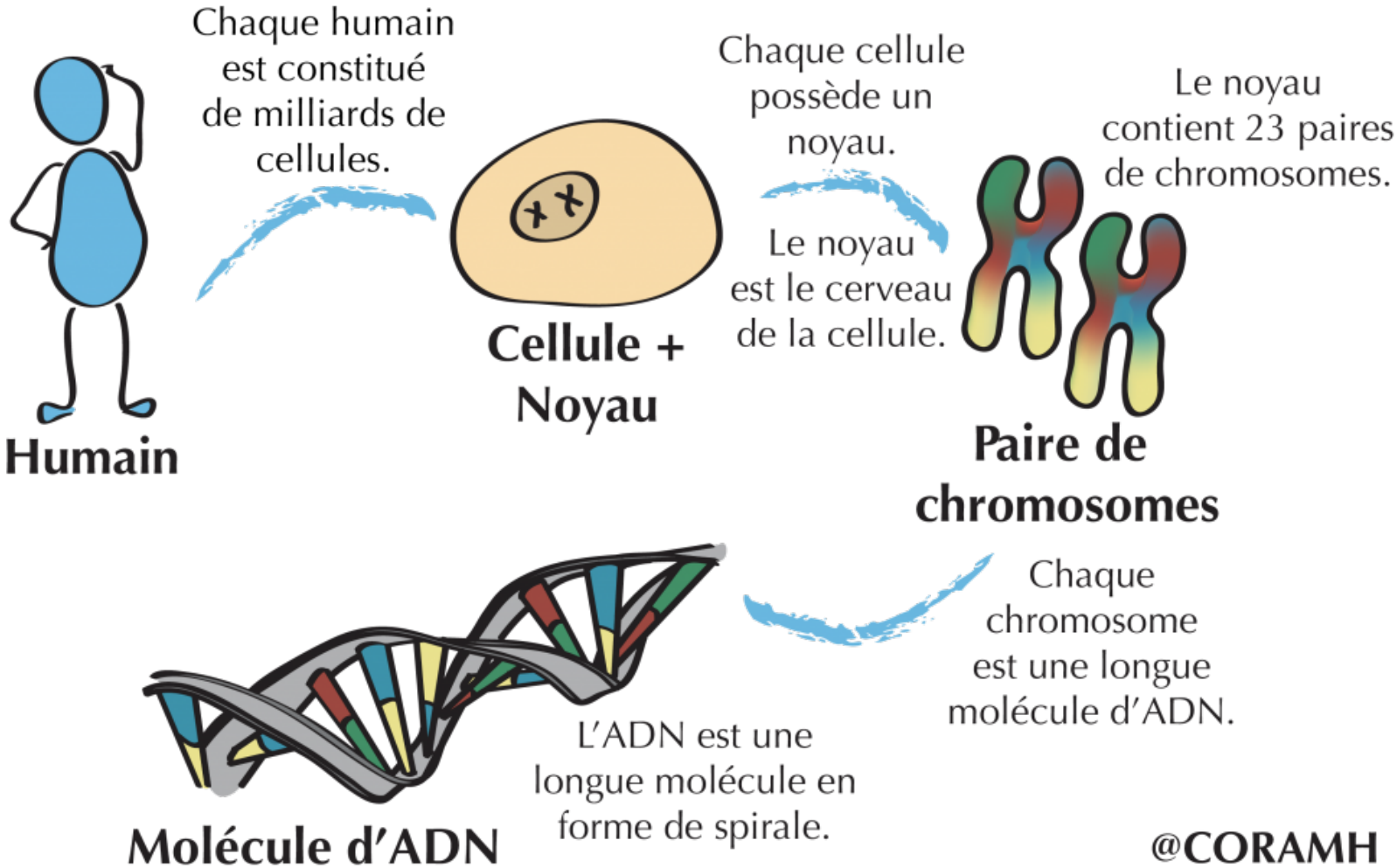
[1] The elements of statistical learning. Springer series in statistics, 2001.

Comment organiseriez-vous une réunion de 20 personnes?



Les médecins misent aussi sur la sparsité

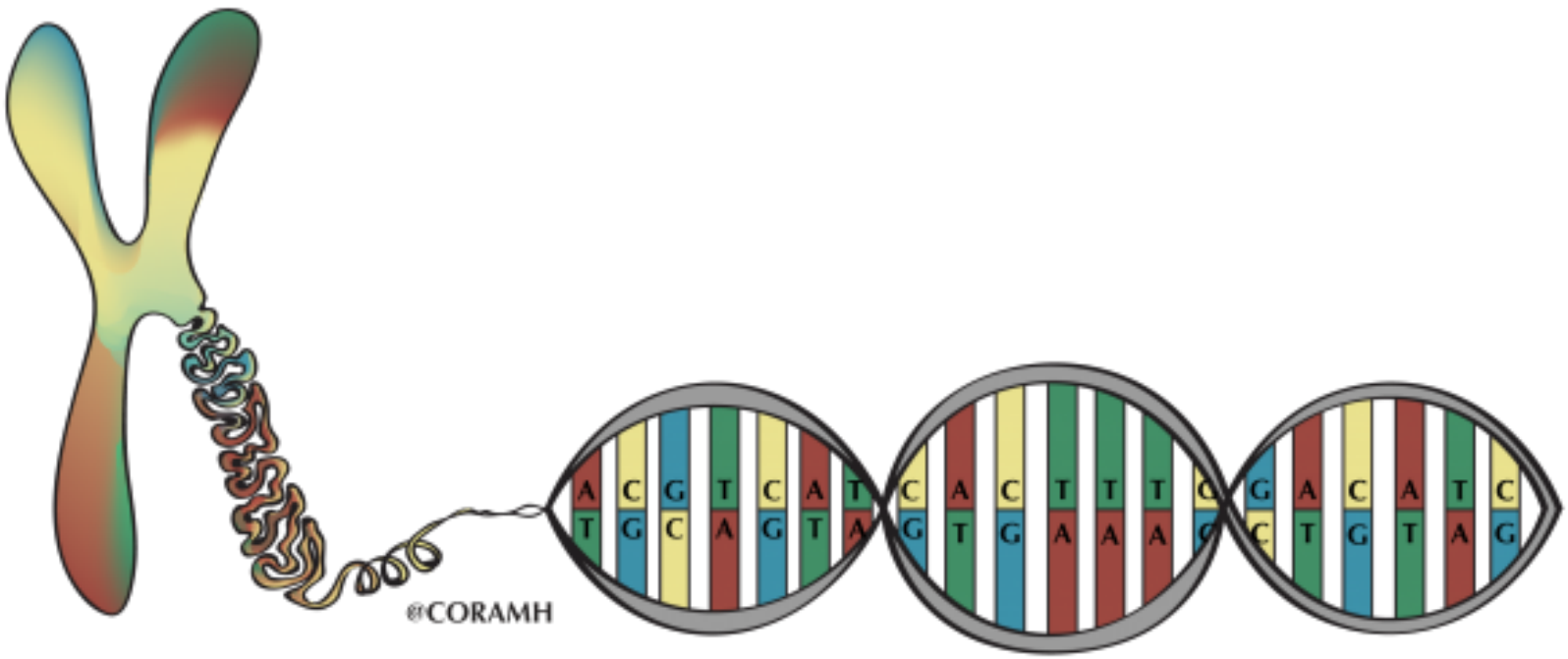
Notions de génétique et d'hérédité



[1] https://coramh.org/genetique-et-heredite/

ADN et gènes

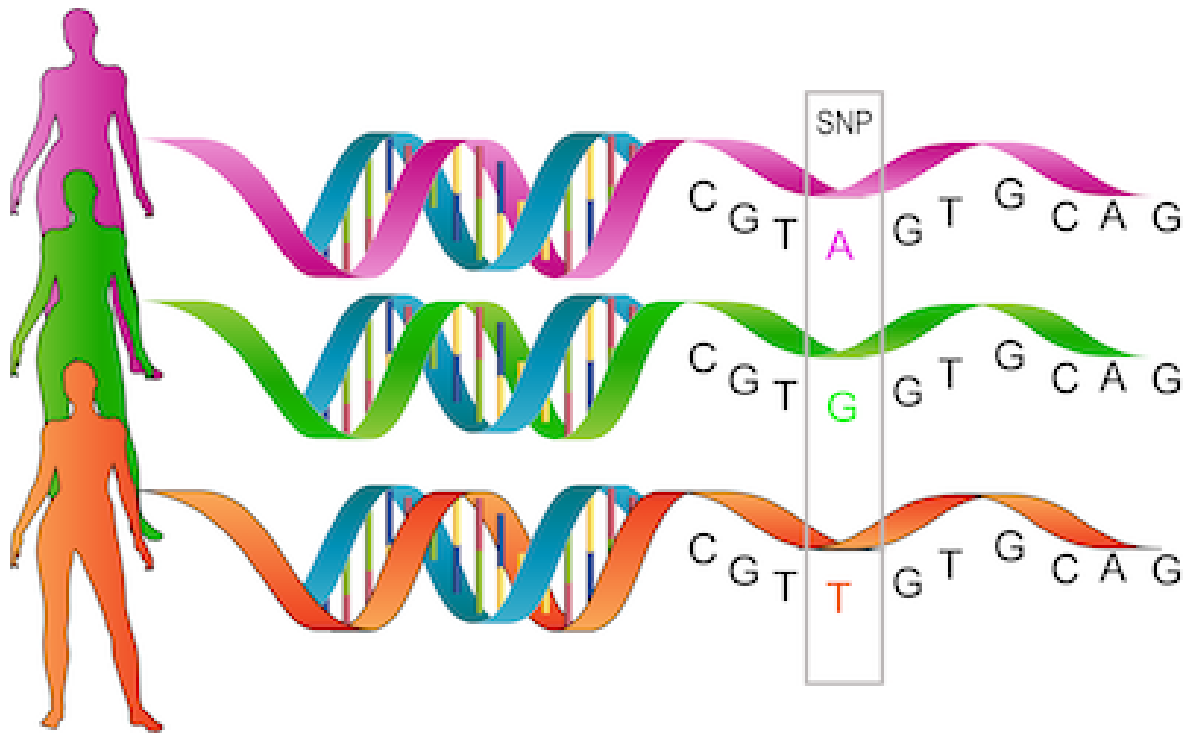
- L'ADN est une molécule contenant l'information génétique écrit dans une langue dont les mots se composent de 4 lettres; A, T, C, G.
- Un gène est un morceau d'ADN formé par la suite précise de plusieurs de ces lettres, cette suite de lettres forme la séquence du gène.
- Il existe plus de 25 000 gènes dans le génome humain qui codent pour différentes caractéristiques physiques et contrôlent le fonctionnement de l'organisme et contribuent à l'état de la santé de l'individu à toutes les étapes de sa vie.



[1] https://coramh.org/genetique-et-heredite/

Single-nucleotide polymorphism (SNP)

- Les SNP sont des régions variables du génome
- La variation doit être située à un endroit spécifique du génome et apparaître sur une proportion supérieure à 1% de la population pour être caractérisée comme SNP



[1] https://www.nature.com/scitable/definition/snp-295/

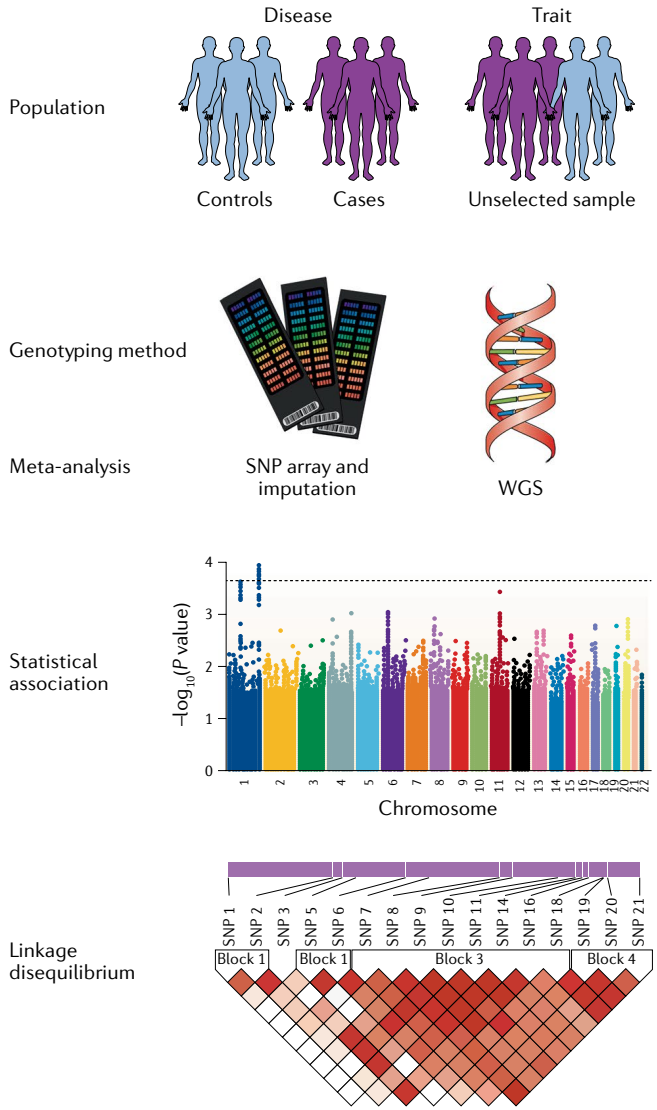
UKBiobank: Données de grande dimension ($n << p$)

- 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

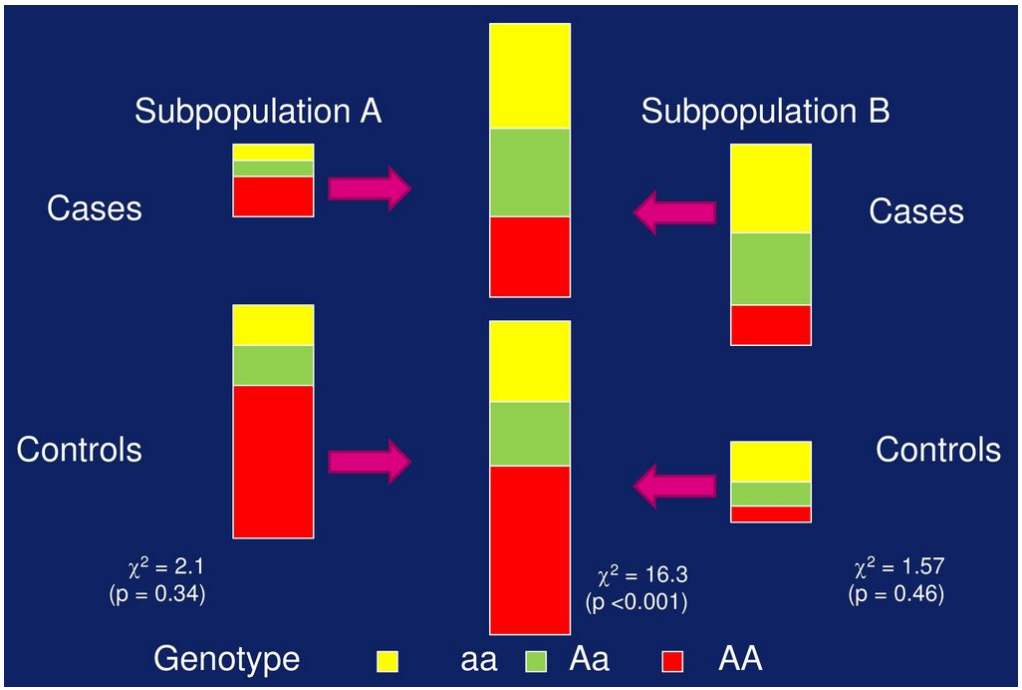
	ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	−1.255	1	2	0	0	0	1
2	4114347	−0.339	1	2	0	2	0	1
3	4399930	−0.6	1	2	1	1	0	1
4	2081319	0.809	1	2	0	1	0	2
5	1347380	0.279	2	2	0	0	0	0
6	3262449	−0.421	2	2	0	1	0	1
7	4870063	−0.454	2	2	0	0	0	2
8	1141212	1.383	2	2	1	1	1	0
9	2997954	−2.29	1	2	0	0	0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

Études d'association pangénomique (GWAS)



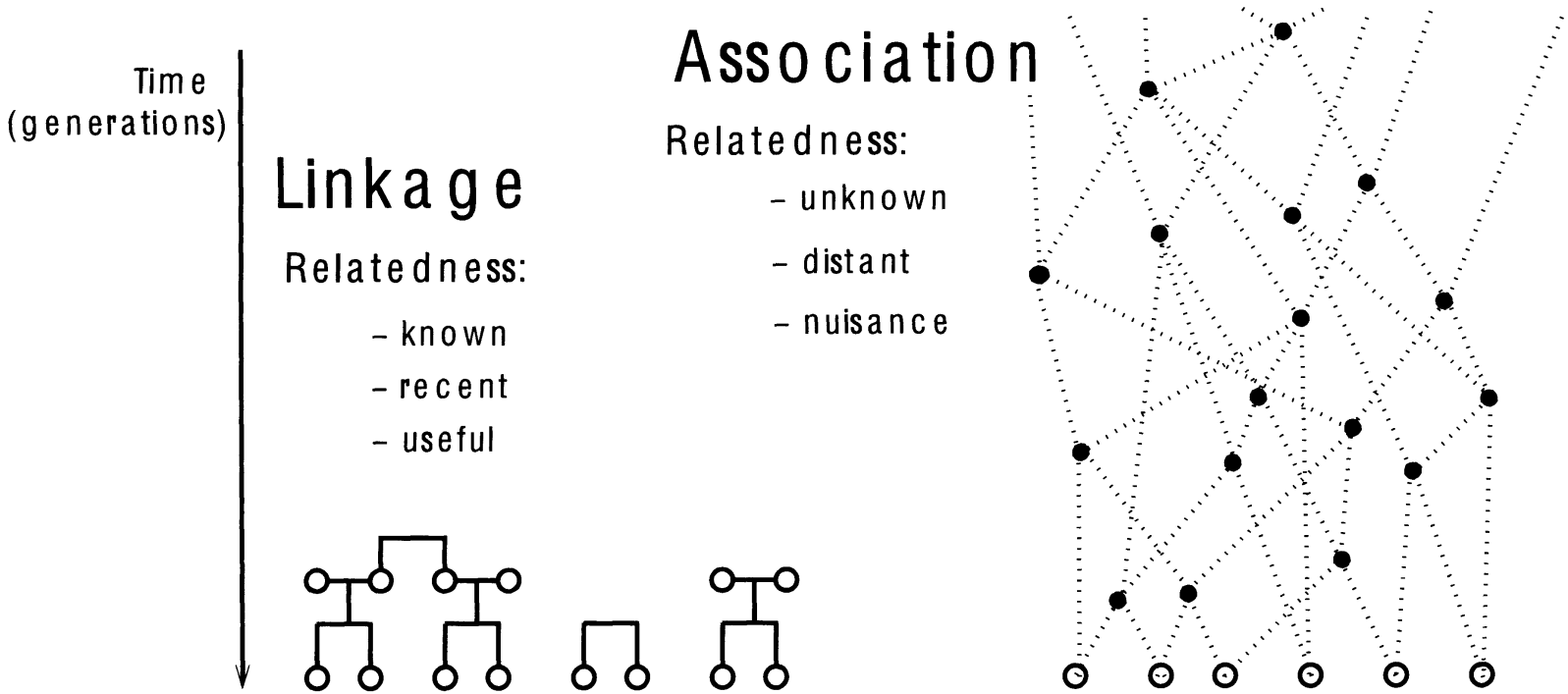
[1] Tam V. et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet (2019)

Facteur de confusion



La structure de population

- 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)

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

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

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

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

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

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

Procédure en deux étapes

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

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

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

- 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

Y

$\sim$

	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

P

+ **E**

Procédure en deux étapes

Step 1:

Y

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

$\sim$

P

	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

$+ E_1$

Step 2:

Residuals from Step 1

$\sim$

	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

$+ E_2$

Procédure en deux étapes

Step 1:

Y

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

$\sim$

P

	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

$+ E_1$

Step 2:

Residuals from Step 1

$\sim$

	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

$+ E_2$

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

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

Check for updates

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

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

Y

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

$\sim$

X

	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

+

P

	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

+ E

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

Data and Model

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

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

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

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

Likelihood

- The negative log-likelihood is given by

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

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

- 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

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

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

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

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

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

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

- 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 ϵ;
for $\lambda \in \{\lambda_{max}, \ldots, \lambda_{min}\}$ **do**
  repeat

$$\text{For } j = 1, \ldots, p, \quad \beta_j^{(k+1)} \leftarrow \arg\min_{\beta_j} Q_\lambda \left(\boldsymbol{\beta}_{-j}^{(k)}, \eta^{(k)}, \sigma^{2\ (k)} \right)$$

$$\eta^{(k+1)} \leftarrow \arg\min_{\eta} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta, \sigma^{2\ (k)} \right)$$

$$\sigma^{2\ (k+1)} \leftarrow \arg\min_{\sigma^2} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta^{(k+1)}, \sigma^2 \right)$$

    $k \leftarrow k + 1$
  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 $\ell(\mathbf{\Theta})$
- 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

[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)

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)

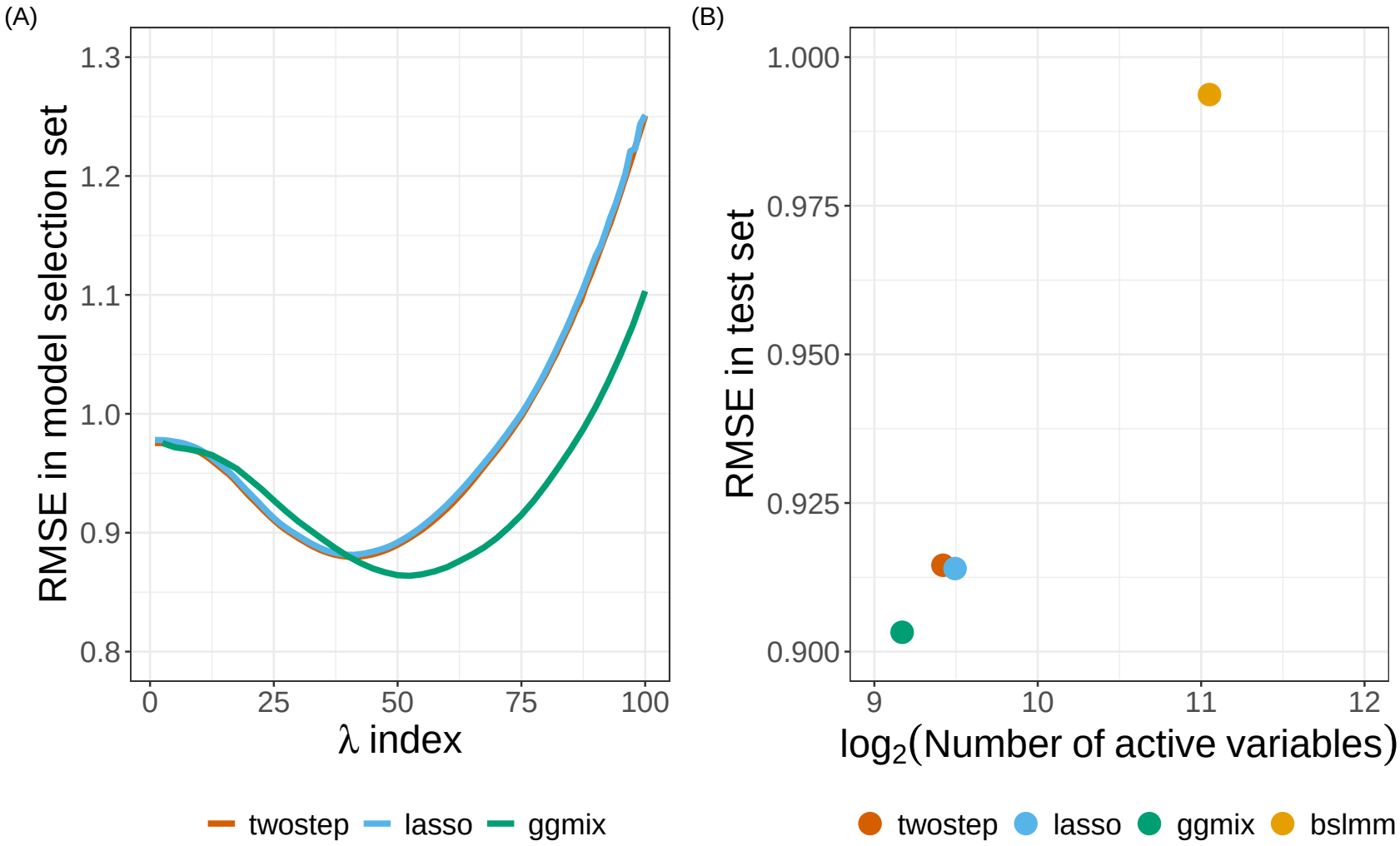
Real data applications

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

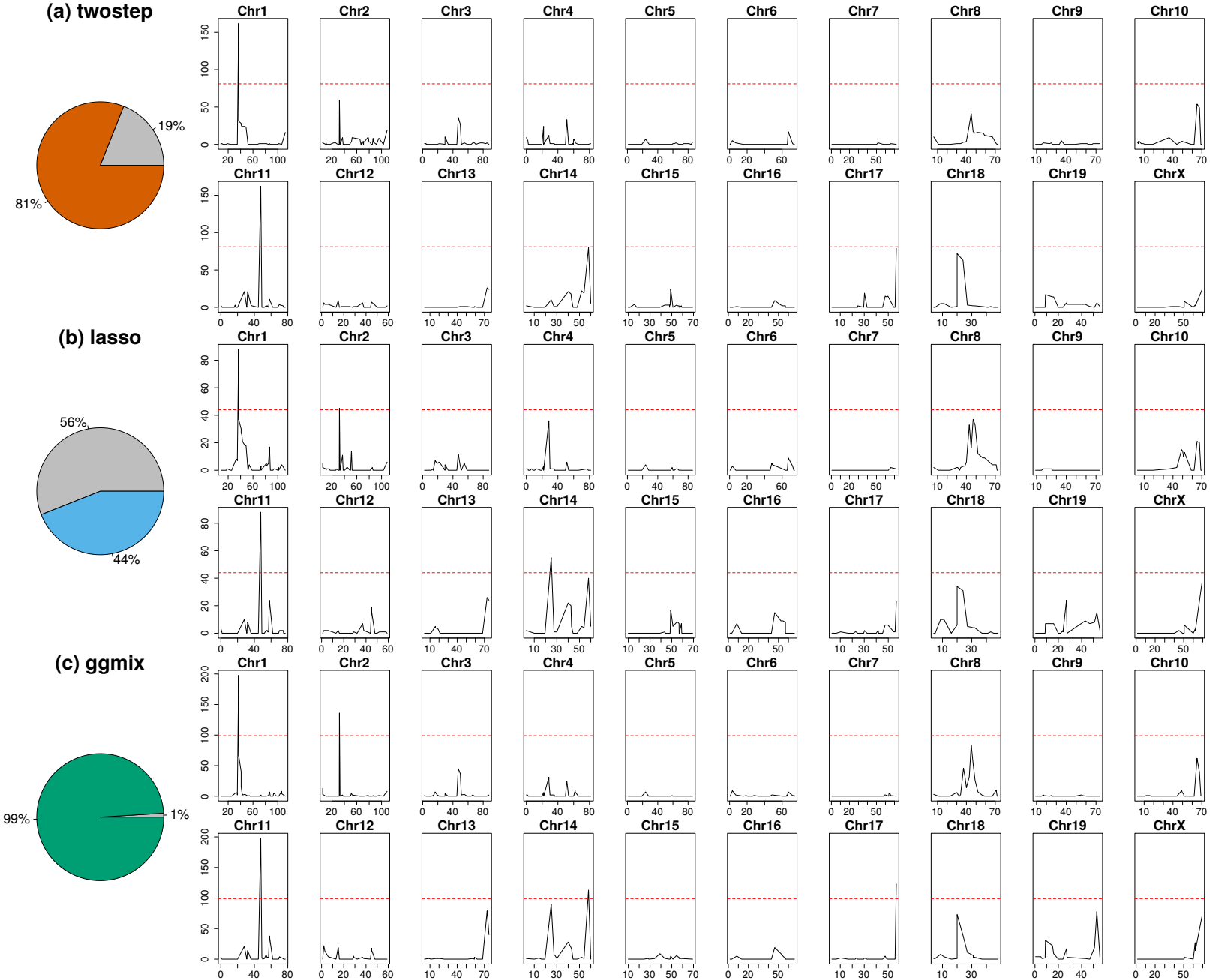


Results: GAW20

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

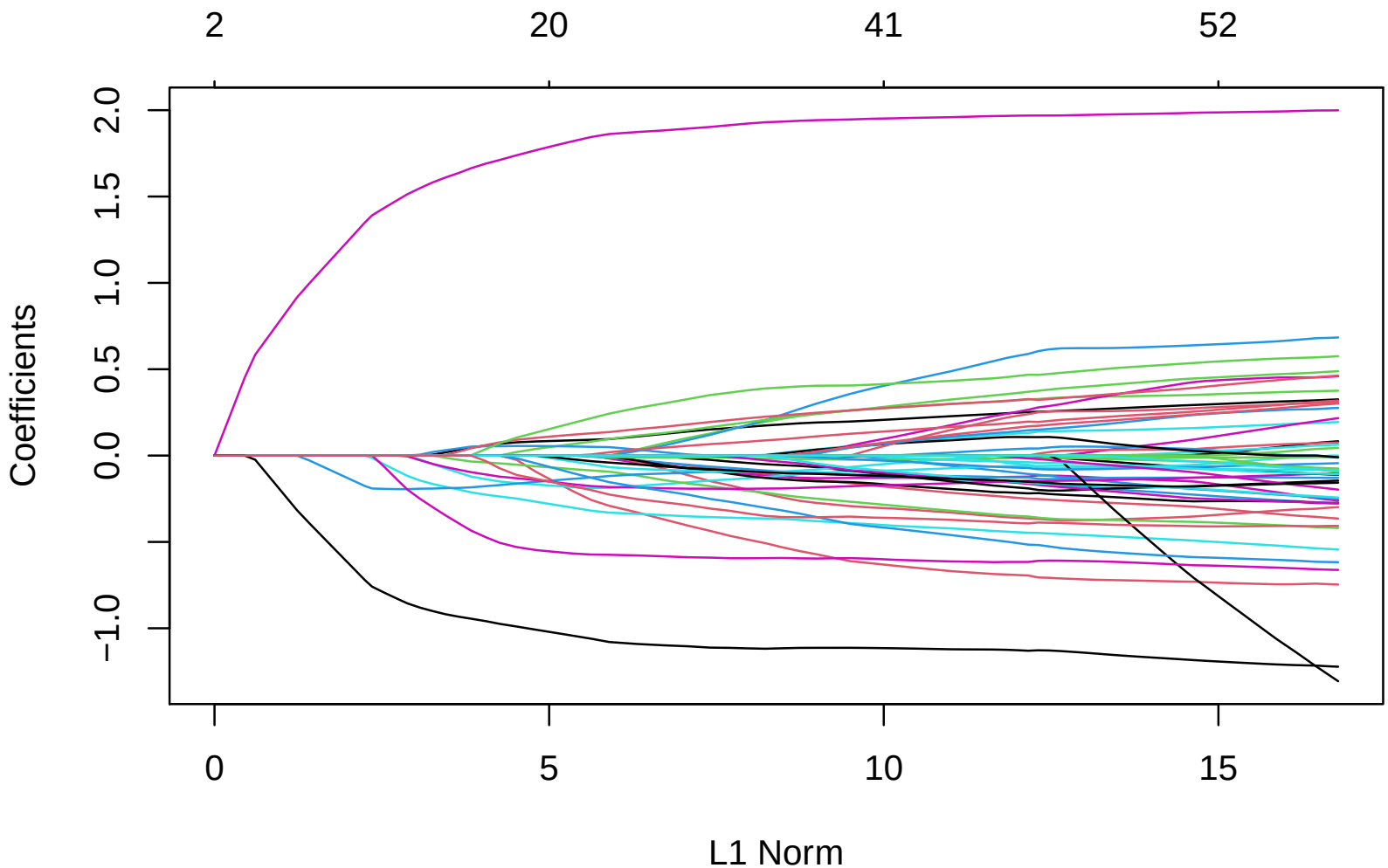


Discussion

- 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)

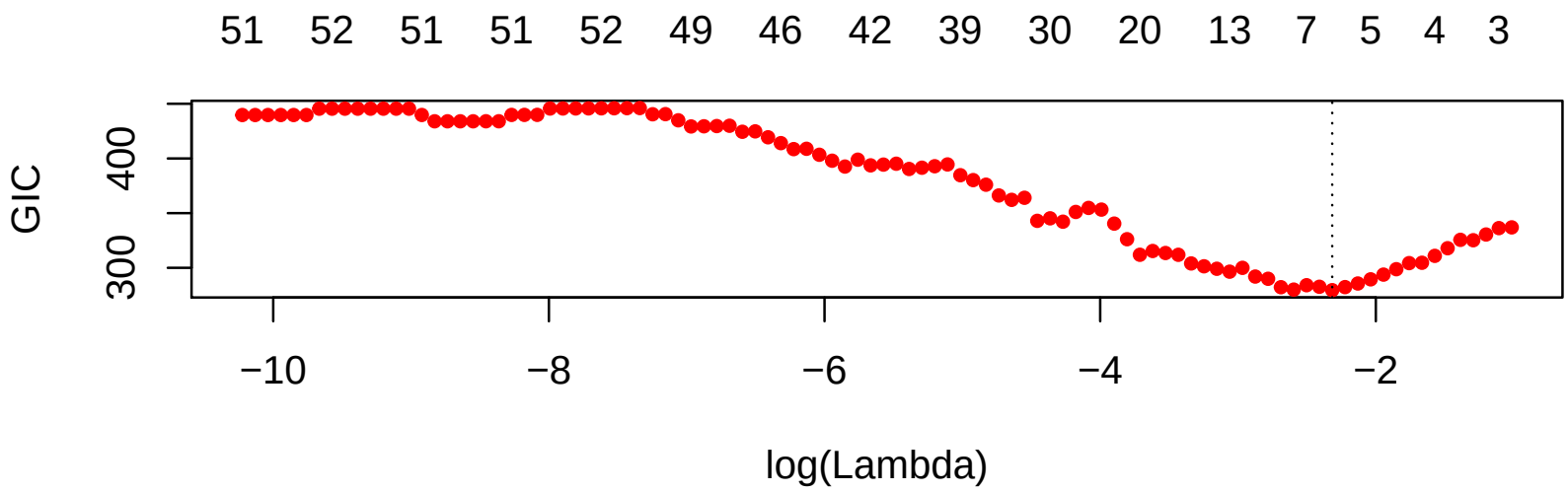
ggmix R package

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



ggmix R package

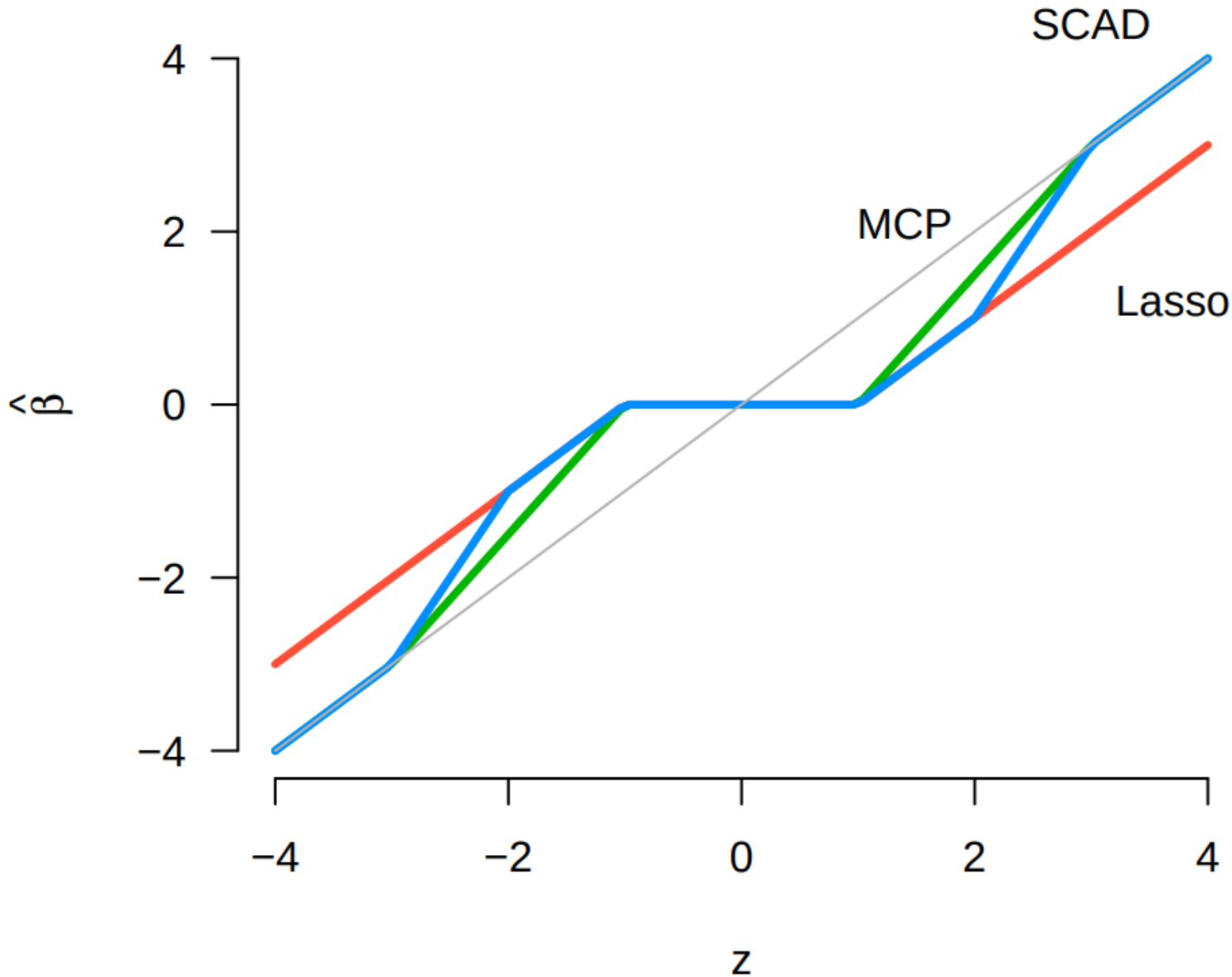
```
hdbic <- gic(fit)
plot(hdbic)
```




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

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

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**

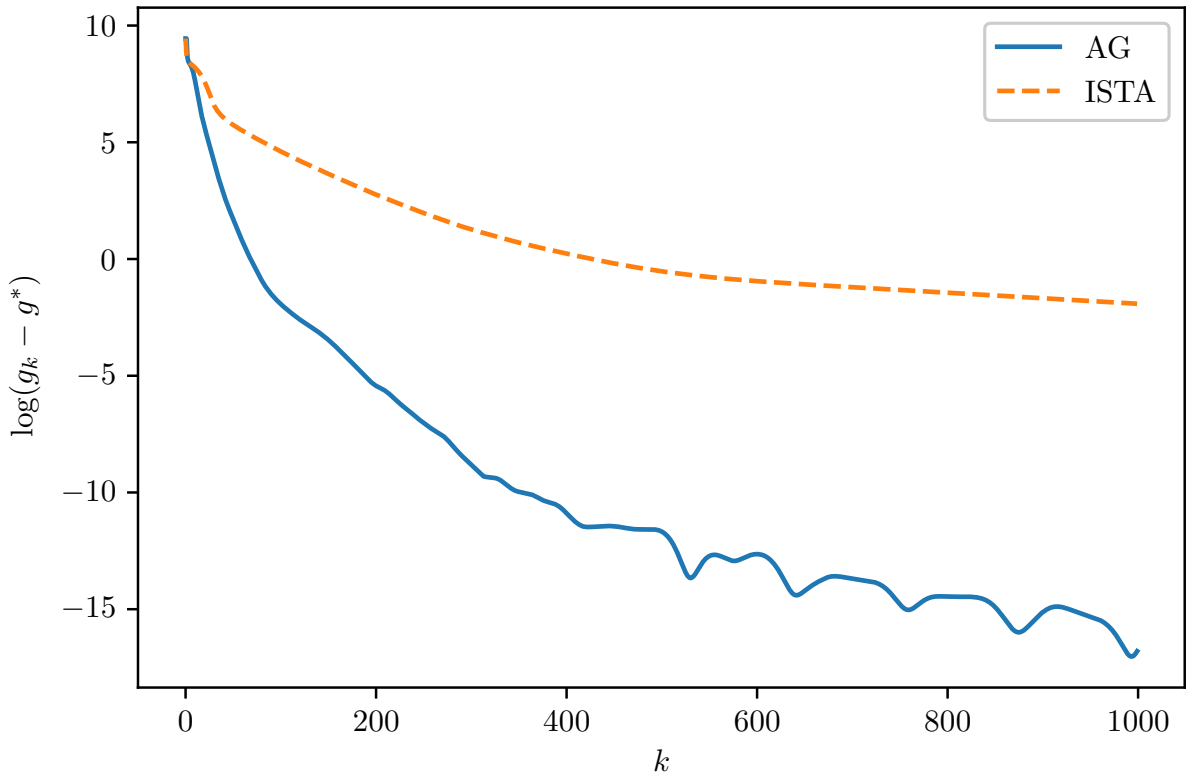
Received: date / Accepted: date

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

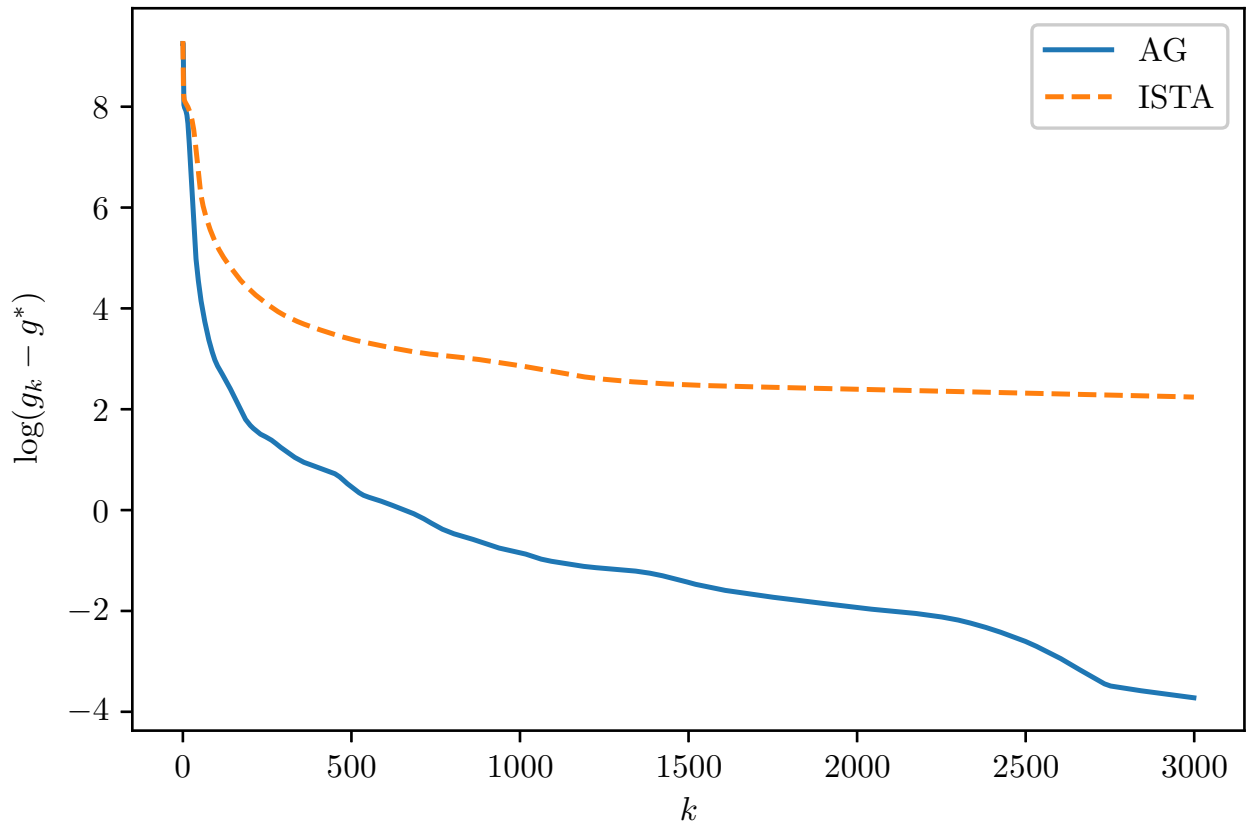
[1] https://arxiv.org/abs/2009.10629

Numerical Study for SCAD

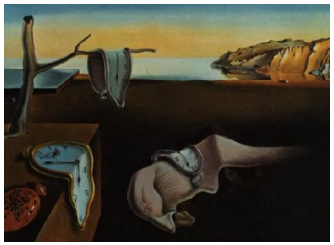


$\mathbf{x}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{I})$, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $\mathbf{y} = \mathbf{X}\boldsymbol{\tau}_{\text{generate}} + \boldsymbol{\varepsilon}$, $\sigma^2 = \frac{\left\|\boldsymbol{\tau}_{\text{generate}}\right\|^2}{3}$, $\boldsymbol{\tau}_{\text{generate}} \in \mathbb{R}^{10006}$ is a sparse constant vector with 6 values of $1.23(\text{intercept}), 3, 4, 5, 6, 59$ 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, $\gamma = 2.5,\ \lambda = 0.6$.



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

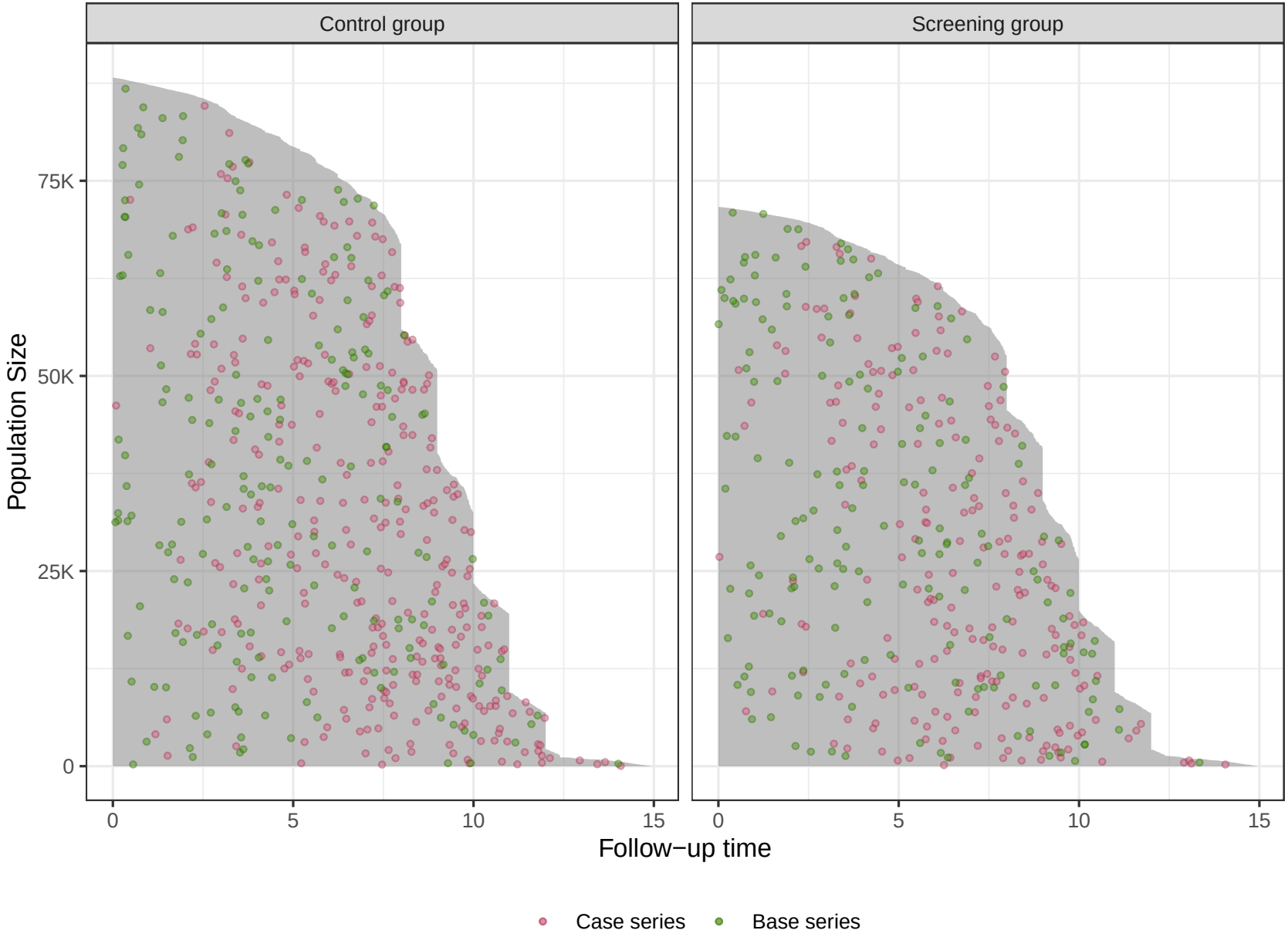
Jesse Islam
McGill University

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: $g(t; \alpha) = \alpha \log t$

Orientations futures

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

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)

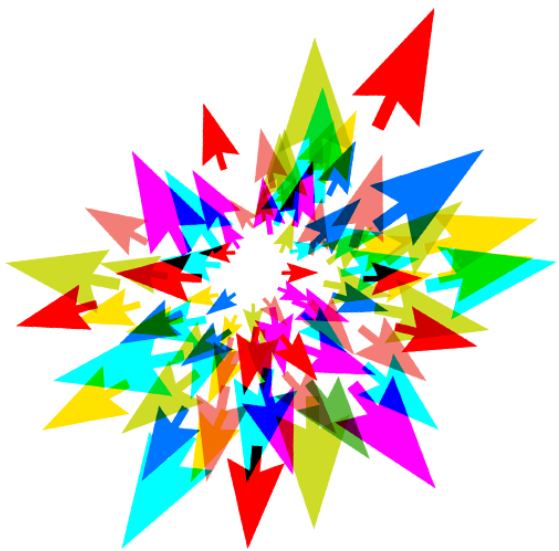
Mohan Zhao, BSc (c)

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



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 20.04 LTS

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

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

other attached packages:
[1] ggmix_0.0.1 knitr_1.30

loaded via a namespace (and not attached):
 [1] lattice_0.20-41  codetools_0.2-16 glmnet_4.0-2     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_0.19
[17] compiler_4.0.2   shape_1.4.5
```