A General Framework for Variable Selection in Linear Mixed Models with Applications to Genetic Studies with Structured Populations

Joint work with

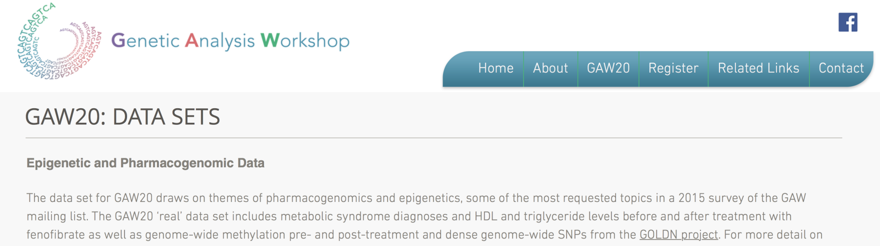
Karim Oualkacha (UQÀM), Yi Yang (McGill), Celia Greenwood (McGill)

JSM, Montreal 2018

sahirbhatnagar.com

Motivation

Genetic Analysis Workshop (GAW20, March 4-7, 2017, San Diego, US)



GAW20: DATA SETS

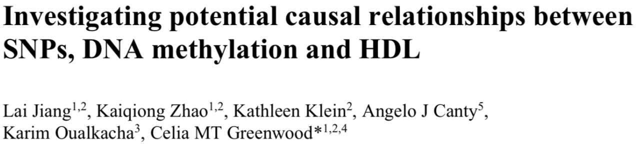
Epigenetic and Pharmacogenomic Data

The data set for GAW20 draws on themes of pharmacogenomics and epigenetics, some of the most requested topics in a 2015 survey of the GAW mailing list. The GAW20 'real' data set includes metabolic syndrome diagnoses and HDL and triglyceride levels before and after treatment with fenofibrate as well as genome-wide methylation pre- and post-treatment and dense genome-wide SNPs from the GOLDN project. For more detail on

[1] GOLDEN project: Genetics of Lipid Lowering Drugs and Diet Network Study

Motivation

- Our contribution in GAW20



Investigating potential causal relationships between SNPs, DNA methylation and HDL

Lai Jiang[1,2], Kaiqiong Zhao[1,2], Kathleen Klein[2], Angelo J Canty[5],
Karim Oualkacha[3], Celia MT Greenwood*[1,2,4]

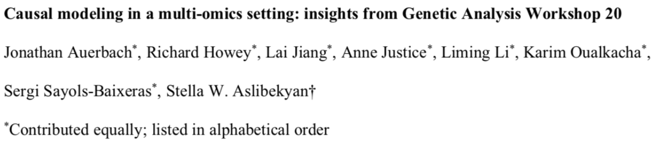
Motivation

- Our contribution in GAW20

Investigating potential causal relationships between SNPs, DNA methylation and HDL

Lai Jiang[1,2], Kaiqiong Zhao[1,2], Kathleen Klein[2], Angelo J Canty[5], Karim Oualkacha[3], Celia MT Greenwood*[1,2,4]

- Contribution of the Causal modelling group



Causal modeling in a multi-omics setting: insights from Genetic Analysis Workshop 20

Jonathan Auerbach*, Richard Howey*, Lai Jiang*, Anne Justice*, Liming Li*, Karim Oualkacha*,

Sergi Sayols-Baixeras*, Stella W. Aslibekyan†

*Contributed equally; listed in alphabetical order

Motivation

- Our contribution in GAW20 consisted of investigation of causal relationship between DNA methylation (exposure) within some genes and ΔHDL (outcome)

Motivation

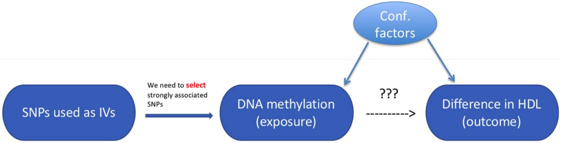
- Our contribution in GAW20 consisted of investigation of causal relationship between DNA methylation (exposure) within some genes and ΔHDL (outcome)
- DNA methylation in these genes has been shown association with HDL

Motivation

- Our contribution in GAW20 consisted of investigation of causal relationship between DNA methylation (exposure) within some genes and ΔHDL (outcome)
- DNA methylation in these genes has been shown association with HDL
- We used Mendelian randomization to explore causal relationship
- We used SNPs around the analyzed genes as Instrumental Variables (IVs) to interrogate the causal relationship

Motivation

- Our contribution in GAW20 consisted of investigation of causal relationship between DNA methylation (exposure) within some genes and ΔHDL (outcome)
- DNA methylation in these genes has been shown association with HDL
- We used Mendelian randomization to explore causal relationship
- We used SNPs around the analyzed genes as Instrumental Variables (IVs) to interrogate the causal relationship



Challenges in GAW20 Data Sets

- GAW20 SNPs data was high-dimensional
- There was a need for data regularization in order to select SNPs strongly associated with the exposure
- Penalized LS regression can be used (Lasso, SCAD, MCP or Elastic net)

Challenges in GAW20 Data Sets

- GAW20 SNPs data was high-dimensional
- There was a need for data regularization in order to select SNPs strongly associated with the exposure
- Penalized LS regression can be used (Lasso, SCAD, MCP or Elastic net)
- But, data consists of families !
- In the GAW20, all regularized methods
  - either did not control for the family structure

Challenges in GAW20 Data Sets

- GAW20 SNPs data was high-dimensional
- There was a need for data regularization in order to select SNPs strongly associated with the exposure
- Penalized LS regression can be used (Lasso, SCAD, MCP or Elastic net)
- But, data consists of families !
- In the GAW20, all regularized methods
  - either did not control for the family structure
  - or used two-steps adjustment for the family structure (including our group)

Challenges in GAW20 Data Sets

- Two-steps adjustment:
  - Step 1 : uses LMM to adjust for subjects relationship

[1] Oualkacha et al. Gene. Epi. (2013)

Challenges in GAW20 Data Sets

- Two-steps adjustment:
  - Step 1 : uses LMM to adjust for subjects relationship
  - Step 2 : uses residuals from Step 1 in variable-selection LS-regression methods to select SNPs

[1] Oualkacha et al. Gene. Epi. (2013)

Challenges in GAW20 Data Sets

- Two-steps adjustment:
  - Step 1 : uses LMM to adjust for subjects relationship
  - Step 2 : uses residuals from Step 1 in variable-selection LS-regression methods to select SNPs
- Two-steps procedure is a valid approach
- In association testing, (GRAMMAR) it is known to suffer from huge power loss [1]

[1] Oualkacha et al. Gene. Epi. (2013)

Proposal

Aim:

We believe that performing variable selection and controlling for familial and/or hidden relationships simultaneously in high-dimensional settings, are likely to be of great interest to the genetic scientists community

Proposal

Aim:

We believe that performing variable selection and controlling for familial and/or hidden relationships simultaneously in high-dimensional settings, are likely to be of great interest to the genetic scientists community

Proposal:

We propose, `ggmix`, a two-in-one procedure which controls for structured populations and performs variable selection in Linear Mixed Models

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: $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$
- 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: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$

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:

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

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

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

Likelihood

- The negative log-likelihood is given by

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

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

Likelihood

- The negative log-likelihood is given by

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

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

- 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

Likelihood

- The negative log-likelihood is given by

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

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

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

$$\boldsymbol{\Phi} = \mathbf{UDU}^\top$$

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

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

Likelihood

- The negative log-likelihood is given by

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

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

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

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

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

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

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: $\tilde{\mathbf{Y}} = \mathbf{U}^\top \mathbf{Y}$
- $\tilde{\mathbf{Y}} | (\beta, \eta, \sigma^2) \sim \mathcal{N}(\tilde{\mathbf{X}}\beta, \sigma^2 \mathbf{W})$, with $\tilde{\mathbf{X}} = \mathbf{U}^\top \mathbf{X}$

Likelihood

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

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

Likelihood

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

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

- For fixed σ^2 and η, solving for β is a weighted least squares problem

Penalized Maximum Likelihood Estimator

- Define the objective function:

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

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

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

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 $\mathbf{\Theta}^{(0)}$ and $\epsilon$;
for $\lambda \in \{\lambda_{max}, \ldots, \lambda_{min}\}$ do
    repeat
        For $j = 1, \ldots, p,\ \beta_j^{(k+1)} \leftarrow \underset{\beta_j}{\arg\min}\ Q_\lambda\left(\beta_{-j}^{(k)}, \eta^{(k)}, \sigma^{2\ (k)}\right)$
                          $\eta^{(k+1)} \leftarrow \underset{\eta}{\arg\min}\ Q_\lambda\left(\beta^{(k+1)}, \eta, \sigma^{2\ (k)}\right)$
                          $\sigma^{2\ (k+1)} \leftarrow \underset{\sigma^2}{\arg\min}\ Q_\lambda\left(\beta^{(k+1)}, \eta^{(k+1)}, \sigma^2\right)$
        $k \leftarrow k + 1$
    until convergence criterion is satisfied:
     $||\mathbf{\Theta}^{(k+1)} - \mathbf{\Theta}^{(k)}||_2 < \epsilon$;
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:

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

- $\widehat{df}_\lambda$ is the number of non-zero elements in $\widehat{\beta}_\lambda$ 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

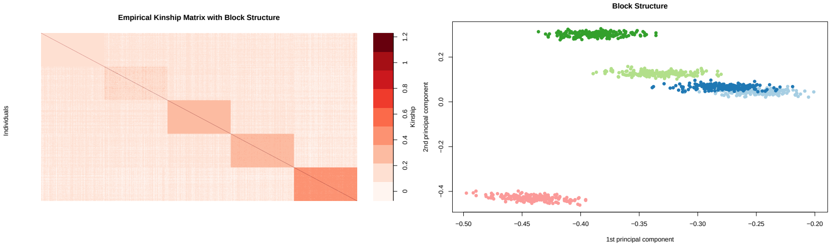
- We simulate genotypes from the BN-PSD Admixture Model[1]
- a : percentage of causal SNPs
- $\mathbf{X}^{(test)}$: $n \times 5000$ matrix of SNPs randomly sampled across the genome
- $\mathbf{X}^{(causal)}$: $n \times (a * 5000)$ matrix of SNPs that are truly associated with the simulated phenotype, $\mathbf{X}^{(causal)} \subseteq \mathbf{X}^{(test)}$
- β_j: effect size for the j^{th} SNP, simulated from a $Uniform(0.3, 0.7)$ for $j = 1, \ldots, (a * 5000)$
- $\mathbf{Y} | (\beta, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}^{(causal)} \beta, \eta\sigma^2 \mathbf{\Phi} + (1 - \eta)\sigma^2 \boldsymbol{\mathcal{I}})$

[1] `https://cran.r-project.org/package=bnpsd`

RRM/Kinship matrix construction

- $\mathbf{X}^{(other)}$: $n \times 10,000$ matrix of simulated SNPs
- $\mathbf{X}^{(kinship)}$: matrix of SNPs used to construct the RRM/Kinship matrix
  - Scenario 1: $\mathbf{X}^{(kinship)} = \mathbf{X}^{(other)} \leftarrow$ No overlap
  - Scenario 2: $\mathbf{X}^{(kinship)} = [\mathbf{X}^{(other)}, \mathbf{X}^{(causal)}] \leftarrow$ 100% overlap
- In each scenario we considered $a = 0, 0.01$, $\eta = 0.1, 0.5$ and $\sigma^2 = 1$

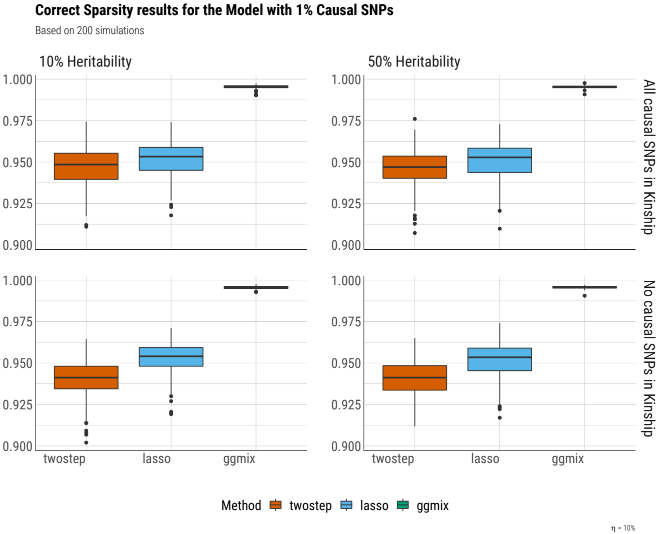
Empirical Kinship Matrix



(a) Kinship Matrix

(b) 1st and 2nd PC

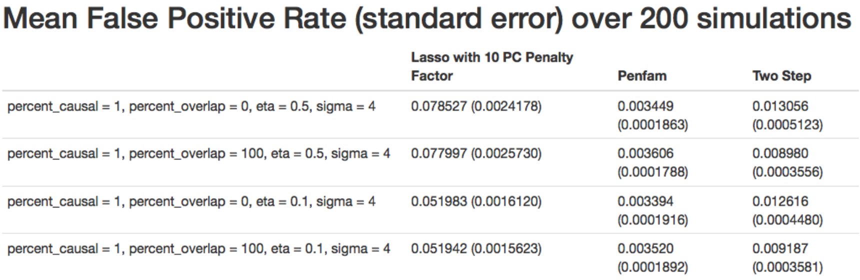
Simulation results



Correct Sparsity results for the Model with 1% Causal SNPs

Based on 200 simulations

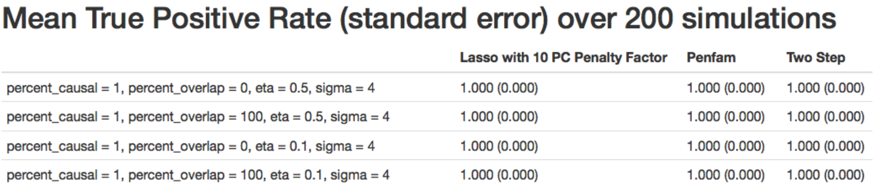
Simulation results



Mean False Positive Rate (standard error) over 200 simulations

	Lasso with 10 PC Penalty Factor	Penfam	Two Step
percent_causal = 1, percent_overlap = 0, eta = 0.5, sigma = 4	0.078527 (0.0024178)	0.003449 (0.0001863)	0.013056 (0.0005123)
percent_causal = 1, percent_overlap = 100, eta = 0.5, sigma = 4	0.077997 (0.0025730)	0.003606 (0.0001788)	0.008980 (0.0003556)
percent_causal = 1, percent_overlap = 0, eta = 0.1, sigma = 4	0.051983 (0.0016120)	0.003394 (0.0001916)	0.012616 (0.0004480)
percent_causal = 1, percent_overlap = 100, eta = 0.1, sigma = 4	0.051942 (0.0015623)	0.003520 (0.0001892)	0.009187 (0.0003581)

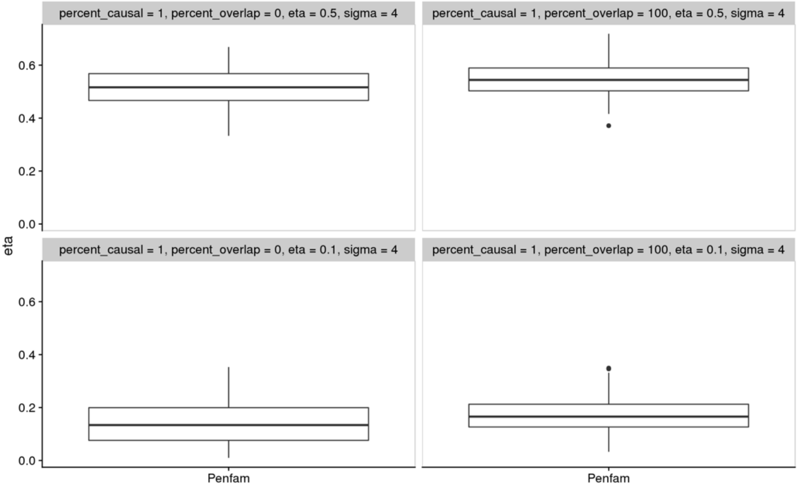
Simulation results



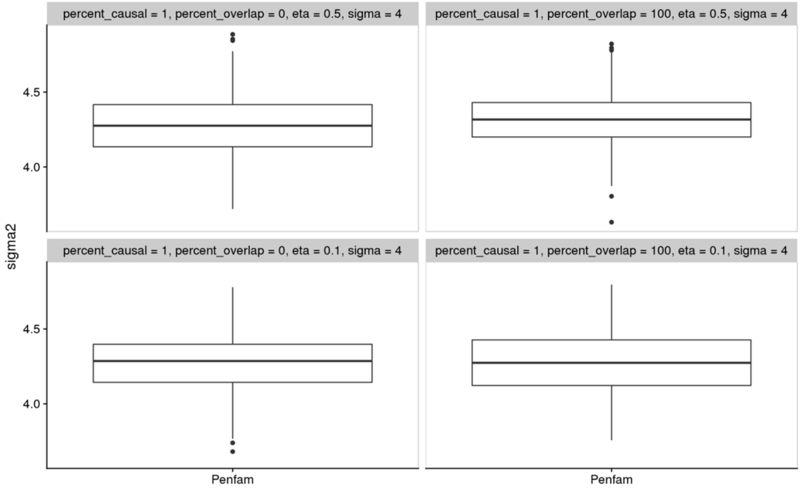
Mean True Positive Rate (standard error) over 200 simulations

	Lasso with 10 PC Penalty Factor	Penfam	Two Step
percent_causal = 1, percent_overlap = 0, eta = 0.5, sigma = 4	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
percent_causal = 1, percent_overlap = 100, eta = 0.5, sigma = 4	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
percent_causal = 1, percent_overlap = 0, eta = 0.1, sigma = 4	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
percent_causal = 1, percent_overlap = 100, eta = 0.1, sigma = 4	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Simulation results



Simulation results



Discussion/Future work

- In some situations, prior information of the predictors (e.g. SNPs) groups structure is available
- Theoretical development of group-Lasso in LMM is already done

Discussion/Future work

- In some situations, prior information of the predictors (e.g. SNPs) groups structure is available
- Theoretical development of group-Lasso in LMM is already done
- In situations where the RRM matrix is of low rank (i.e. $n >> \#$ of SNPs used to construct RRM). ex: UK Biobank
- Computational time of fitting `ggmix` can be reduced using SVD decomposition of $\mathbf{X}^{(kinship)}$ in order to construct $\mathbf{\Phi}$ and in order to transforme the data
- Theoretical development of low-rank trick is already done

Discussion/Future work

- Capturing the subjects relationship using random effect requires VCs estimation
- Random effect modelling leads to a non-convex optimization problem
- Fixed effects models are good alternatives to random effects models for analysis of Longitudinal/Panel data [1]
- Capturing familial structure using a penalized FE model could be an interesting avenue to explore

[1] Roger Koenker, JMA, (2004)