Heritability estimation in high dimensional mixed models

Anna Bonnet

Supervisors: Elisabeth Gassiat and Céline Lévy-Leduc

Journées MAS Grenoble

29 août 2016

Heritability

- Heritability of a biological trait: Proportion of phenotypic variance explained by genetic factors.
- Estimation of heritability in human genetics: better understanding of complex diseases, further research for genetic causes...
- Estimation of heritability in animal and vegetal genetics: determination of optimal genotypes to produce a valuable resource.

Examples of data sets - Quantitative traits

- Vector of observations : $Y = \begin{pmatrix} 162 \\ 181 \\ \dots \\ 175 \end{pmatrix}$

- Predictors : $X = \begin{pmatrix} 17 \\ 32 \\ \dots \\ 25 \end{pmatrix}$

- Matrix of SNPs : $W = \begin{pmatrix} 0 & 1 & \dots & 1 \\ 0 & 2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 2 \end{pmatrix}$

Framework of genetic studies, $n \sim 2000$ individuals, $N \sim 500000$ SNPs

Sparse Linear Mixed Model

$$Y = X\beta + Zu + e$$

where
- Y is a vector n $\times$ 1 of observations
- $X\beta$ are the fixed effects
- Z is a random matrix n $\times$ N, centered and normalized version of W.
- u and e are the random effects

$$u_i \overset{i.i.d.}{\sim} (1-q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2}) \text{ , for all i and } e \sim \mathcal{N}\left(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right)$$

$\triangleright$ Estimation of $\eta^\star = \frac{Nq\sigma_u^{\star 2}}{Nq\sigma_u^{\star 2}+\sigma_e^{\star 2}}$.

Heritability estimator

Up to considering the projection of Y onto $(\text{Im } X)^{\perp}$, we focus on the model

$$Y = Zu + e$$

- In the case $q = 1$ (no sparsity),

$$Y|Z \sim \mathcal{N}\left(0, \eta^\star \sigma^{\star 2} ZZ'/N + (1-\eta^\star)\sigma^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right).$$

- $\hat{\eta}$ is defined as the maximizer of the log-likelihood conditionally to Z :

$$L_n(\eta) = -\log\left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\eta(\lambda_i - 1)+1}\right) - \frac{1}{n}\sum_{i=1}^n \log\left(\eta(\lambda_i - 1) + 1\right)$$

where $\widetilde{Y} = U'Y$ and $U\frac{ZZ'}{N}U' = diag(\lambda_1, ..., \lambda_n)$.

- Method implemented in the **R package HiLMM**.

Theoretical result

Theorem

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ *satisfy the sparse LMM with* $\eta^\star > 0$ *and assume that the random variables* $Z_{i,j}$ *are i.i.d.* $\mathcal{N}(0,1)$.
Then for any $q \in (0,1]$, *as* $n, N \to \infty$ *such that* $n/N \to a > 0$,

$$\sqrt{n}(\hat{\eta} - \eta^\star)$$

converges in distribution to a centered Gaussian random variable with variance

$$\tau^2(a, \eta^\star, q) = \frac{2}{\widetilde{\sigma}^2(a, \eta^\star)} + 3\frac{a^2 \eta^{\star 2}}{\widetilde{\sigma}^4(a, \eta^\star)} \left(\frac{1}{q} - 1\right) S(a, \eta^\star)$$

where $\widetilde{\sigma}^2(a, \eta^\star)$ *and* $S(a, \eta^\star)$ *are positive functions, for which closed-form expressions are available.*

Simulation study

Influence of sparsity q Influence of $a = n/N$

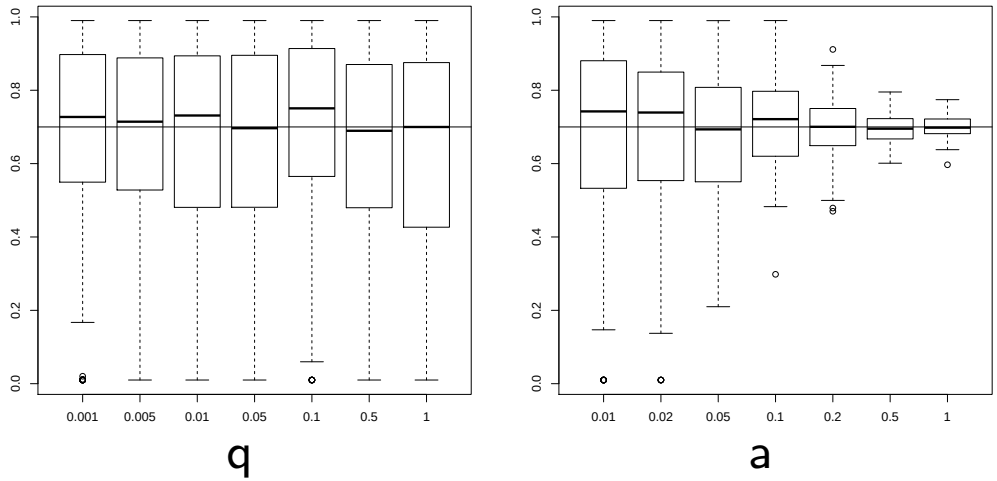


Figure: Boxplots of $\hat{\eta}$ for different values of q when $a = 0.01$ (right) and different values of $a = \frac{n}{N}$ when $q = 1$(left).

▷ When a decreases, that is $N >> n$, the variance of our heritability estimator increases.
▷ The presence of null components ($q < 1$) does not influence the estimations.

Variable selection steps

- **Step 1: Empirical correlation computation (SIS, Fan & Lv (2008))** . It consists in reducing the number of relevant columns of Z by trying to remove those associated to null components in the vector u. The matrix reduced to the most significant columns is denoted Z_{red}.

- **Step 2: The LASSO criterion**. It consists in minimizing with respect to u the following criterion:

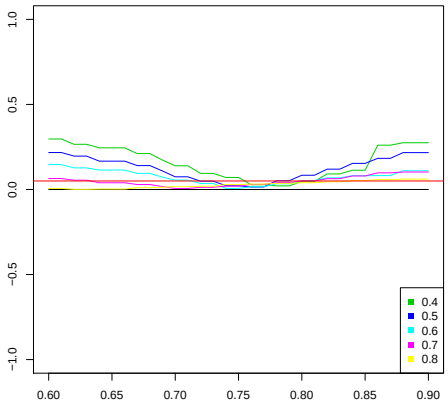
$$Crit_\lambda(u) = \|Y - Z_{red}u\|_2^2 + \lambda\|u\|_1$$

The choice of λ is made according to the **stability selection** method (Meinshausen, 2010).

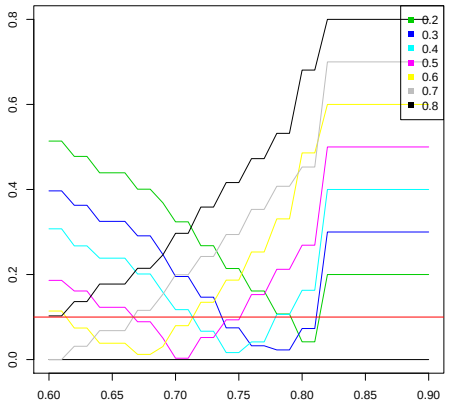
- **R Package EstHer**: Variable selection + Heritability Estimation + Computation of standard errors

Choice of the threshold in the stability selection step

▷ Each choice of threshold gives a set of selected variables, and then an estimated value of the heritability.



100 causal SNPs



10000 causal SNPs

Figure: Absolute difference $|\eta^\star - \hat{\eta}|$ for thresholds from 0.6 to 0.9 and for 100 (left) and 10000 (right) causal SNPs.

▷ For 100 causal SNPs, there is a range of thresholds between 0.7 and 0.85 which provide a good estimation for heritability, with 0.78 as optimal threshold.
▷ For 10000 causal SNPs, there does not exist such a threshold.

First results of the variable selection method

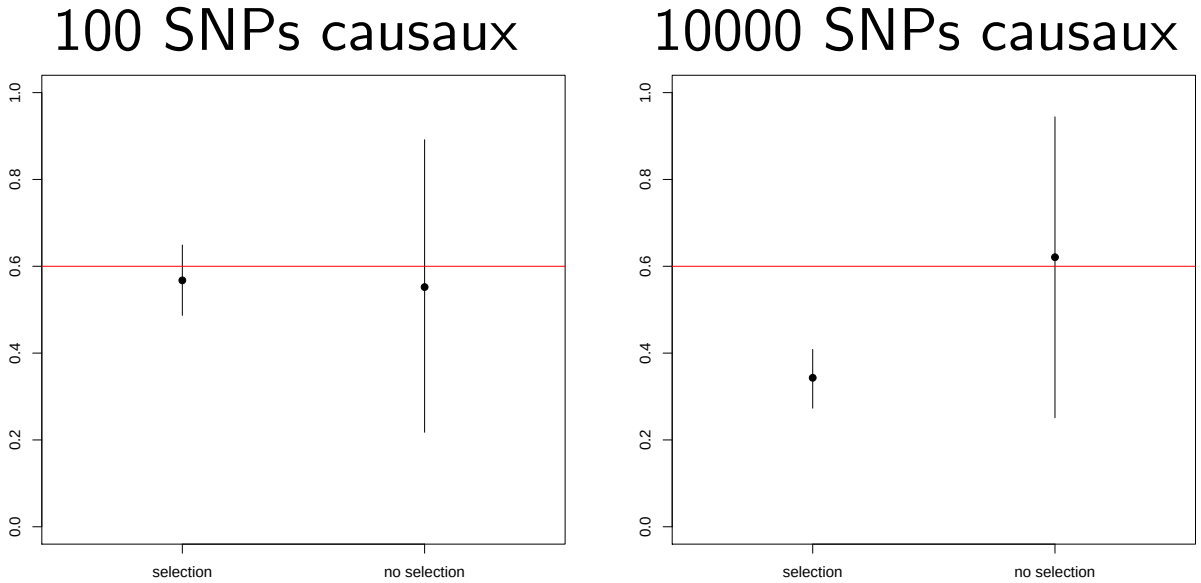


Figure: Estimation of $\eta^\star$ using our variable selection method with threshold 0.78 and using no variable selection.

▷ For 100 causal SNPs, selecting variables reduces substantially the variance.
▷ For 10000 causal SNPs, selecting variables creates an important bias.

Results for different thresholds

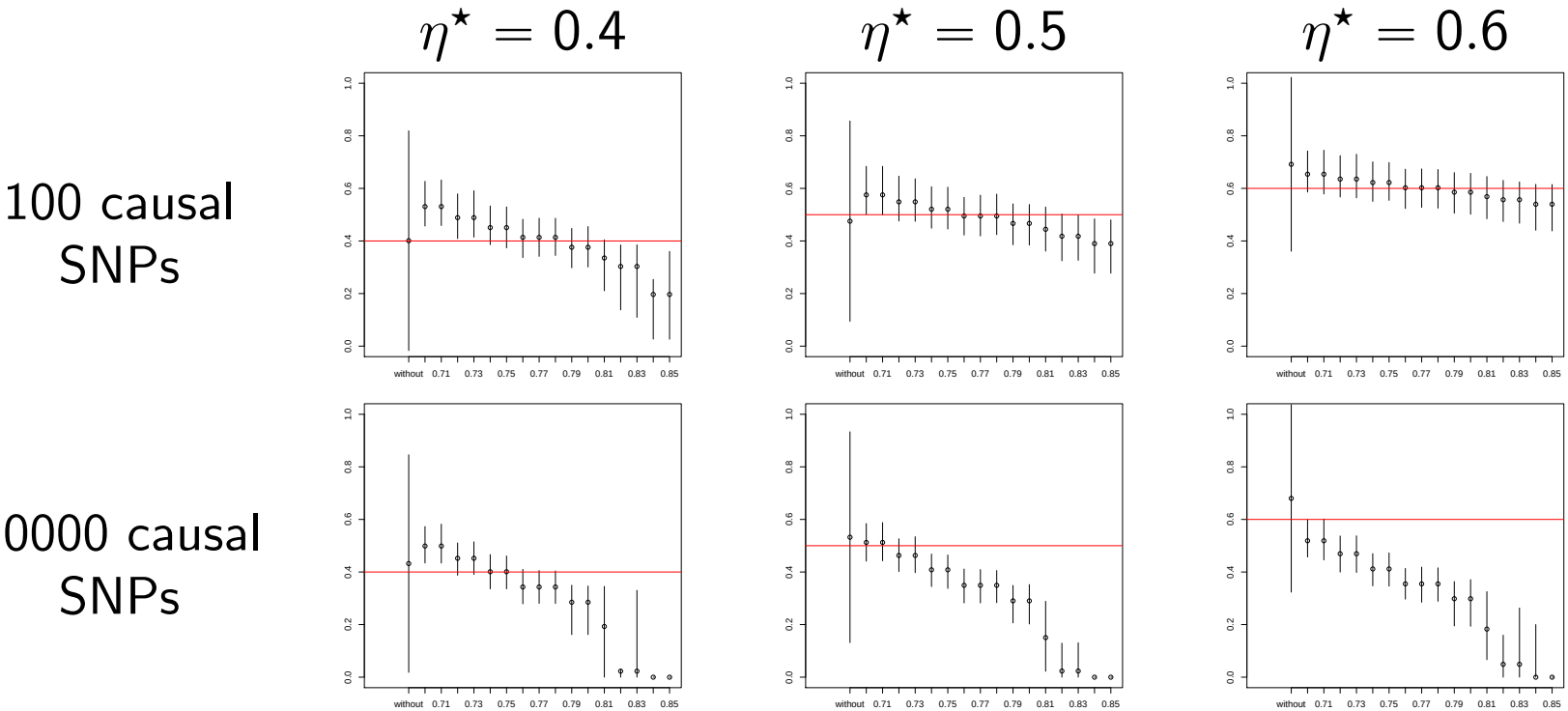


Figure: Estimation of the heritability with 95% confidence intervals obtained without selection and with selection and for thresholds between 0.7 and 0.85.

▷ 100 causal SNPs: two close thresholds provide similar estimations.
▷ 10000 causal SNPs: a small change in the threshold causes substantial differences in the estimations.

A criterion to decide whether to apply the variable selection or not

Table: Mean value of the number (and proportion) of overlapping confidence intervals for 16 thresholds from 0.7 to 0.85.

$\eta^\star$	100 causal SNPs	1000 causal SNPs	10000 causal SNPs
0.4	12.2 (0.76)	6.6 (0.41)	6.9 (0.43)
0.5	14.9 (0.93)	6.6 (0.41)	6.3 (0.39)
0.6	16 (1)	7.8 (0.48)	7.2 (0.45)

$\triangleright$ Criterion: If the mean proportion of overlapping thresholds is greater than 0.6, we perform variable selection with threshold 0.78, otherwhise we estimate directly the heritability.

Application of the criterion

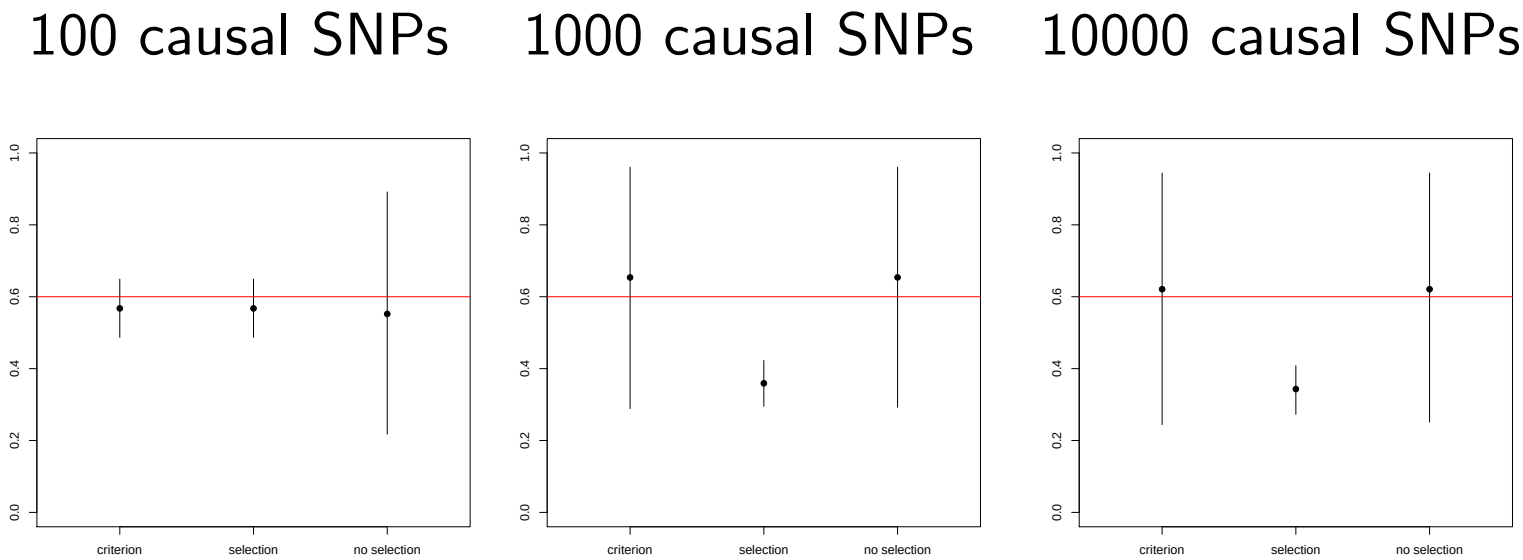


Figure: Comparison of our method with the criterion, the methods with and without selection.

▻ Introducing the criterion allows our estimator to have a smaller variance than the estimator without selection when the number of causal SNPs is small, and to have the same behavior when the number of causal SNPs is high.

Application to brain volume data

Data from the project Imagen: volume of the different regions of the brain from ∼2000 adolescents in Europe.

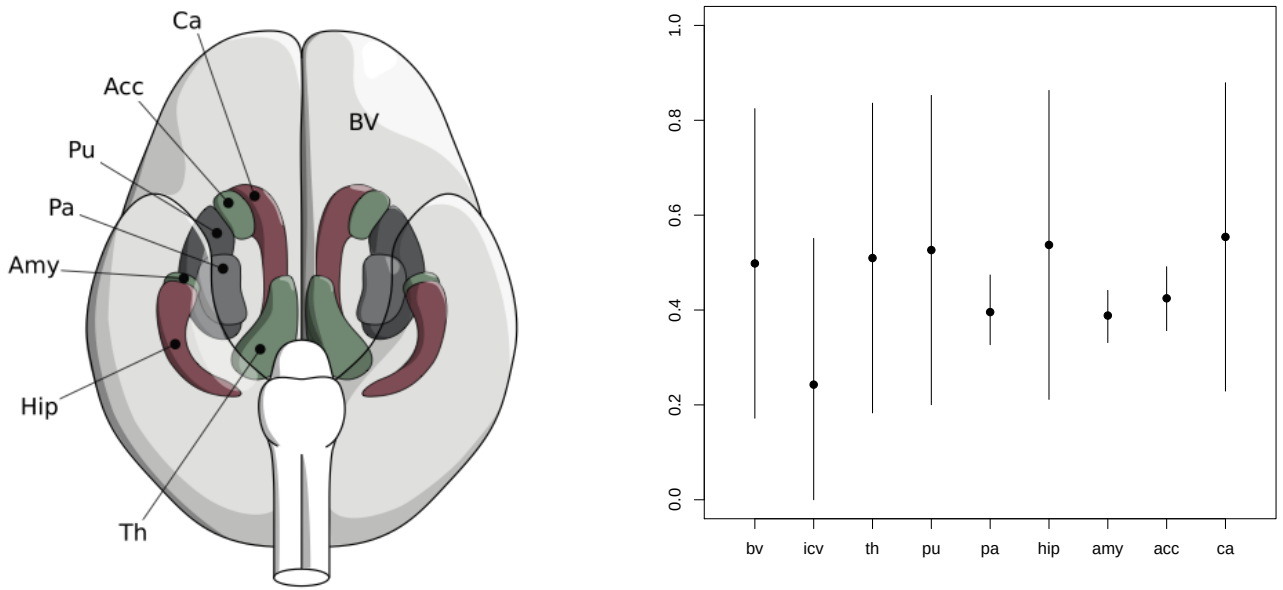


Figure: Different regions of the brain (Toro et al, 2014) and the estimation of heritability for these different regions' volumes.

Extension to binary data

- How to define heritability for binary traits?

Liability model (Falconer, 1965)

$$\mathbf{Y}_i = \mathbb{1}_{\{\mathbf{L}_i > t\}}$$

where

$$\mathbf{L} = \mathbf{Z}\mathbf{u} + \mathbf{e},$$

with $\mathbf{L} = (\mathbf{L}_1, \ldots, \mathbf{L}_n)$, $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2} I_N)$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2} I_n)$

- The heritability is defined "at the liability scale", that is

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}.$$

Case-control studies

- Specificity of case-control studies: the cases are highly oversampled. The number of patients and controls are similar even for rare diseases.
- Least square method (Golan, 2014) which takes into account this oversampling of the cases:

$$\hat{\eta} = \underset{\eta \in (0,1)}{argmin} \sum_{i \neq j} (p_i p_j - \mathbb{E}[p_i p_j | \mathbf{Z}, S = 1])^2$$

- $p_i = \frac{Y_i - P}{\sqrt{P(1-P)}}$
- p prevalence in the study
- $\{S = 1\}$ if individuals i and j are in the study.

$\rightarrow$ Approximation of $\mathbb{E}[p_i p_j | \mathbf{Z}, S = 1]$.

Approach

$$\mathbb{E}(p_i p_j|\mathbf{Z}, S=1) = \frac{1-P}{P}\mathbb{P}(Y_i = Y_j = 1|\mathbf{Z}, S=1) - \mathbb{P}(Y_i \neq \mathbf{Y}_j|\mathbf{Z}, S=1) + \frac{P}{1-P}\mathbb{P}(Y_i = Y_j = 0|\mathbf{Z}, S=1).$$

- Approximation of $\mathbb{P}(Y_i = Y_j = 1|\mathbf{Z})$, $\mathbb{P}(Y_i = Y_j = 0|\mathbf{Z})$, $\mathbb{P}(Y_i \neq Y_j|\mathbf{Z})$.

$$\mathbb{P}(Y_i = Y_j = 1|\mathbf{Z}) = \int_t^\infty \int_t^\infty f(x,y)dxdy,$$

$$\text{where } f(x,y) = \frac{1}{2\pi}|\Sigma^{(N)}|^{-\frac{1}{2}} \exp\left\{-\frac{(x,y)\Sigma^{(N)-1}(x,y)^t}{2}\right\}.$$

$$\text{with } \Sigma^{(N)} = \begin{pmatrix} 1+\eta^\star \frac{B_i}{\sqrt{N}} & \eta^\star \frac{C_{i,j}}{\sqrt{N}} \\ \eta^\star \frac{C_{i,j}}{\sqrt{N}} & 1+\eta^\star \frac{B_j}{\sqrt{N}} \end{pmatrix}$$

where $B_i = O_p(1)$, $B_j = O_p(1)$ and $C_{i,j} = O_p(1)$.

Approximation and corresponding estimator

- First order approximation:

$$\mathbb{E}(p_i p_j | Z, S = 1) = c G_{i,j} \eta^\star$$

where

- $G_{i,j} = \frac{1}{N} \sum_{i=1}^{N} Z_{i,k} Z_{j,k}$

- c a constant which depends on the prevalence K in the population, the prevalence P in the study and the threshold t.

- The heritability estimator has an explicit form

$$\hat{\eta} = \frac{\sum_{i \neq j} p_i p_j G_{i,j}}{\sum_{i \neq j} G_{i,j}^2}$$

Consistency of the heritability estimator

Theorem (Consistency)

$\hat{\eta}$ is a consistent estimator of $\eta^\star$, that is

$$\hat{\eta} \xrightarrow{P} \eta^\star$$

when $n \to +\infty$, $N \to +\infty$ and $n/N \to a > 0$, under mild assumptions on the matrix Z.

Numerical results

- Comparison of the estimators $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$ obtained respectively with the first and second order approximations of $\mathbb{E}[p_i p_j | \mathbf{Z}, S_i = S_j = 1]$.

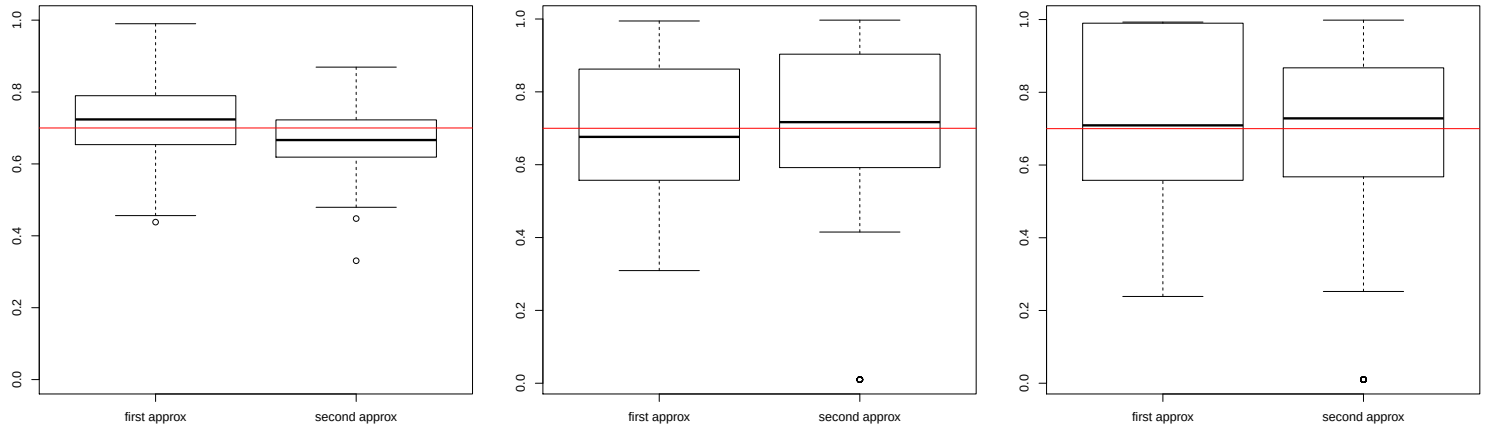


Figure: Performance of $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$ for $n = 100$, $N = 10000$ and different values of k: 0.1 (left), 0.01 (middle) and 0.005 (right).

▷ The numerical results obtained with the two approximations are similar.

Conclusions and perspectives

- Conclusions

  - Quantitative traits: we proposed a hybrid estimator which includes a selection step in very sparse scenarios and behaves like the maximum likelihood estimator otherwhise.

  - Binary traits: we showed the consistency of the heritability estimator proposed by Golan et al. (2014).

- Perspectives

  - Quantitative traits: study the biological pathways between the lists of selected SNPs.

  - Binary traits: consider sparsity, build accurate confidence intervals.

References

[1] Anna Bonnet, Elisabeth Gassiat, and Celine Levy-Leduc. Heritability estimation in high-dimensional sparse linear mixed models. *Electronic Journal of Statistics*, 9(2):2099–2129, 2015.

[2] Anna Bonnet, Elisabeth Gassiat, Celine Levy-Leduc, Roberto Toro, and Thomas Bourgeron. Improving heritability estimation by a variable selection approach in sparse high dimensional linear mixed models, 2016. Submitted.

[3] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[4] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[5] Nicolai Meinshausen and Peter Buhlmann. Stability selection. *Journal of the Royal Statistical Society*, pages 417–473, 2010.