Heritability estimation in high-dimensional mixed models

Anna Bonnet, Elisabeth Gassiat, Céline Lévy-Leduc

Banff - March 28, 2017

Heritability

- Heritability of a biological trait: Proportion of phenotypic variance explained by genetic factors.

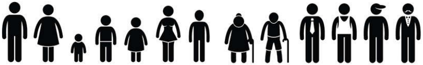
Phenotype (P) = Genotype (G) + Environment (E)

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

Heritability: $H^2 = \frac{\sigma_G^2}{\sigma_P^2}$

Heritability

- The biological trait can be either quantitative or qualitative.



Quantitative (Height)

Binary (Disease)

- Interest of estimating heritability: better understanding of complex diseases, further research for genetic causes...

Linear Mixed Model

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

where

- Y is a $n \times 1$ vector of observations
- $X\beta$ are the fixed effects (age, city, ...)
- Z is a $n \times N$ random matrix which contains the genetic information (SNPs matrix)
- u and e are independent random effects

$$u \sim \mathcal{N}(0, \sigma_u^{\star 2} \mathrm{Id}_{\mathbb{R}^N}) \text{ and } e \sim \mathcal{N}\left(0, \sigma_e^{\star 2} \mathrm{Id}_{\mathbb{R}^n}\right)$$

- Classical mathematical definition of heritability :

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

Sparse Linear Mixed Model

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

where

- Y is a $n \times 1$ vector of observations
- $X\beta$ are the fixed effects
- Z is a $n \times N$ random matrix, which contains the genetic information
- u and e are the random effects

$$u_i \overset{i.i.d.}{\sim} (1 - q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2})\ , \text{for all } i$$

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

Heritability estimator

In the sequel, we consider

$$Y = Zu + e$$

- We study the maximum likelihood estimator in the case $q = 1$ (no sparsity): misspecification of the model.
- Reparameterization with new parameters $\eta^\star$ and $\sigma^{\star 2} = N\sigma_u^{\star 2} + \sigma_e^{\star 2}$ (Pirinen et al. 2013).

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

- $\hat{\eta}$ maximizer of the log-likelihood conditionally to Z.

Framework

Our methodology is inspired from Yang et al. (2011) and Pirinen et al. (2013) but the theoretical properties of this estimator have not been established.

- State of the art: $q = 1$, N is fixed and $n \to \infty$.
- In genetic applications, $N >> n$, q is unknown.
- Our goal: establish theoretical properties about our estimator in the framework $q \in (0, 1]$, $n, N \to \infty$ and $n/N \to a \in (0, +\infty)$.

$\sqrt{n}$-Consistency

Theorem

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ *satisfy the sparse LMM with* $\eta^\star > 0$ *and* $\hat{\eta}$ *the maximizer of* $L_n(\eta)$.
Then, under mild assumptions on Z, *for all* q *in* $(0, 1]$, *as* $n, N \to \infty$ *such that* $n/N \to a \in (0, +\infty)$,

$$\sqrt{n}(\hat{\eta} - \eta^\star) = O_P(1).$$

Central Limit Theorem in the sparse LMM

Theorem

Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$ satisfy the sparse LMM with $\eta^\star > 0$ and assume that $Z_{i,j}$ 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

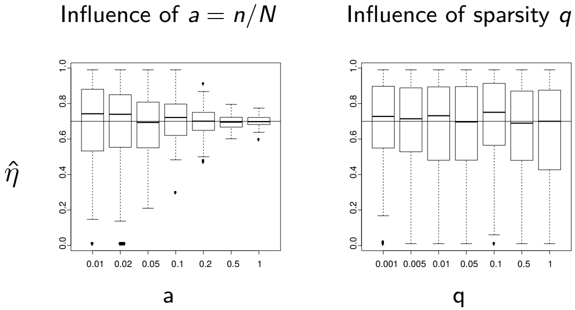


Figure: Estimations of $\eta^\star$ for $n = 1000$ and for different values of $a = \frac{n}{N}$ when $q = 1$ (left) and different values of q when $a = 0.01$ (right).

- 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

- **Step 1: Empirical correlation computation (SIS, Fan & Lv (2008))** . We keep the columns of Z which are the most correlated to Y. The reduced matrix is denoted Z_{red}.

- **Step 2: The LASSO criterion**. We minimize with respect to u the criterion:

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

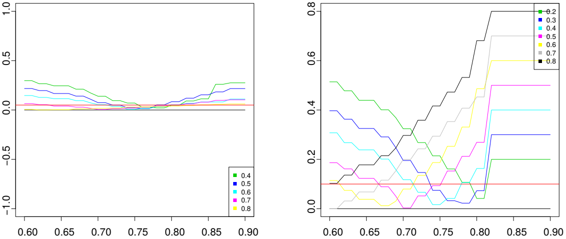
$+$ **stability selection** (Meinshausen & Buhlmann, 2010).

- **Step 3: Bootstrap method** to compute confidence intervals.

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

Choice of the threshold in the stability selection step

- A choice of threshold → a set of selected variables, an estimated value of $\eta^\star$



100 causal SNPs

10000 causal SNPs

Figure: Absolute difference $|\eta^\star - \hat{\eta}|$ for thresholds from 0.6 to 0.9.

- 100 causal SNPs: a range of thresholds (0.7-0.85) provides a good estimation for heritability (optimal threshold: 0.78)
- 10000 causal SNPs: no optimal threshold.

First results of the variable selection method

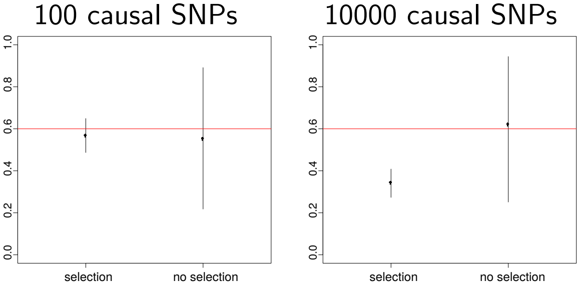


Figure: Estimation of $\eta^\star$ using our variable selection method with threshold 0.78 and using no variable selection ($n = 2000, N = 100000$).

- For 100 causal SNPs, selecting variables reduces substantially the variance.
- For 10000 causal SNPs, selecting variables leads to underestimate $\eta^\star$.

Influence of the threshold in the stability selection

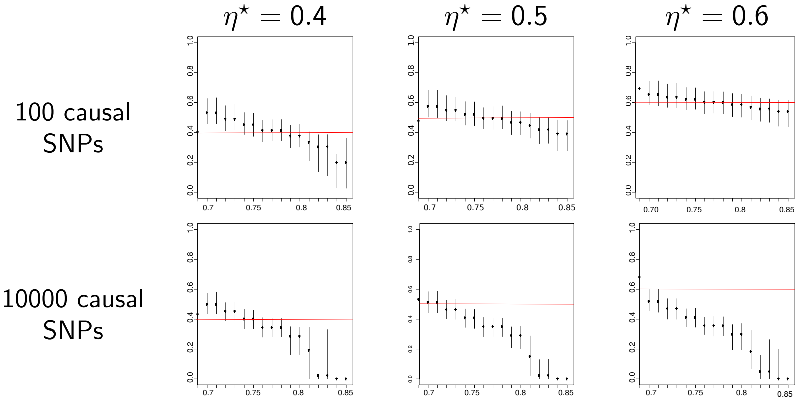


Figure: Heritability estimations with 95% CI for thresholds between 0.7 and 0.85.

- 100 causal SNPs: two close thresholds provide similar estimations.
- 10000 causal SNPs: small change in the threshold → very different estimations.

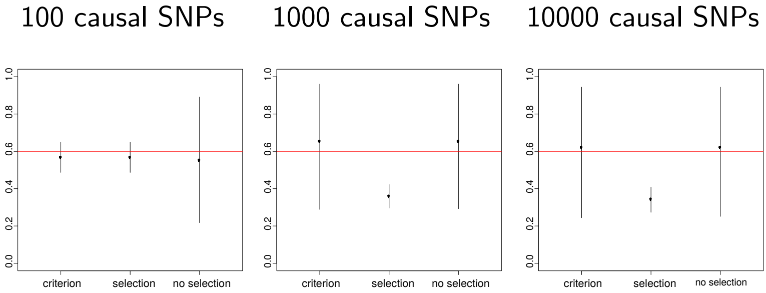
A criterion to decide whether to apply the variable selection or not

Table: Mean value (and proportion) of the number 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)

- **Criterion:** If the mean proportion of overlapping thresholds > 0.6 → variable selection.

Application of the criterion

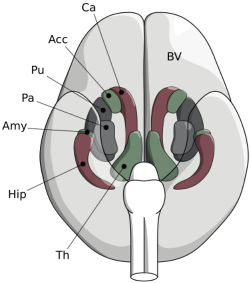


- Small number of causal SNPs: reduction of standard errors
- High number of causal SNPs: behaves like HiLMM (no selection).

Application to brain volume data

Collaboration with T.Bourgeron's GHFC team (Institut Pasteur)

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



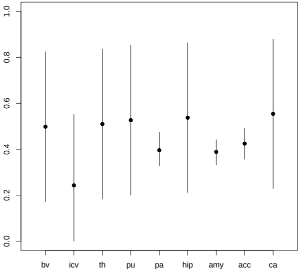


Figure: Different regions of the brain (Toro et al, 2014) and the estimation of heritability for these different regions' volumes.

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.

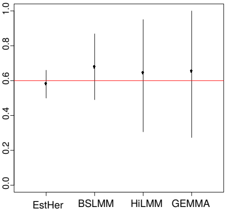
[4] Nicolai Meinshausen and Peter Buhlmann. Stability selection. *Journal of the Royal Statistical Society*, pages 417–473, 2010.

[5] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

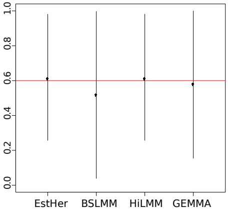
Comparison

- BSLMM (Zhou et al, 2013): Bayesian method which can adapt to sparsity.

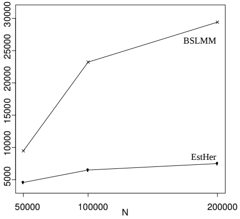
100 causal SNPs



10000 causal SNPs



Computational times (in seconds)



- Convergence issues when using the default parameters in BSLMM.
- EstHer faster than BSLMM.