Robustness and power of single-SNP analysis in populations with related individuals.

Simon Teyssèdre[1] J-M. Elsen[1] A. Ricard[2]

[1] INRA, UMR 1313, 78352 Jouy-en-Josas, France
[2] INRA, UR 631, 31326 Castanet-Tolosan, France

QTL-MAS Workshop 2010, Poznan, Poland
simon.teyssedre@toulouse.inra.fr

Outline

1. Introduction
   - Context
   - Objectives
   - Methods tested
2. Materials & Methods
   - Model and test
   - Model on SNP
3. Results
   - Regression
   - GRAMMAR
   - QTDT
4. Discussion & On going

Outline

1. Introduction
   - Context
   - Objectives
   - Methods tested
2. Materials & Methods
   - Model and test
   - Model on SNP
3. Results
   - Regression
   - GRAMMAR
   - QTDT
4. Discussion & On going

Context

- What we know :
  - Animal populations are often populations with related individuals
  - Conventional methods (LD) make the assumption that there is no relatedness between individuals ⇒ Population structure will affect the robustness of the methods
  - People try to control the population structure in their models

Context

- What we know :
  - Animal populations are often populations with related individuals
  - Conventional methods (LD) make the assumption that there is no relatedness between individuals ⇒ Population structure will affect the robustness of the methods
  - People try to control the population structure in their models
- But we often make the first analysis from simple methods (ex: simple regression) that do not correct for population structure (fast, easy to implement)

Context

- What we know :
  - Animal populations are often populations with related individuals
  - Conventional methods (LD) make the assumption that there is no relatedness between individuals ⇒ Population structure will affect the robustness of the methods
  - People try to control the population structure in their models
- But we often make the first analysis from simple methods (ex: simple regression) that do not correct for population structure (fast, easy to implement)
- Objective : Provide tools to show that some of these first analysis may give erroneous results in populations with related individuals

Objectives

- Objective : Evaluate algebraically the robustness and power of some methods from known parental structure (h^2, pedigree, Nb individuals ...)
- Hypothesis : We assume that the true model is :

$$y = 1\mu + x\beta + Zu + e$$

  with Y the vector of records, μ the overall mean, β the SNP effect, u the random polygenic effect with $u \sim N(0,A\sigma_u^2)$, A the relationship matrix and e the residuals with $e \sim N(0,I\sigma_e^2)$
- Idea : The user, who doesn't know the reality and therefore the true model, is wrong and uses a different model (for example, he uses the same model without the polygenic effect (Zu))
- Question : how many false positives should be expected if we use this simple regression? Will there be sufficient power to detect a QTL?

Methods tested

True model :

$$y = 1\mu + x\beta + Zu + e$$

We tested 3 different methods (all are single SNP analysis) :

1. Simple regression (LD):

$$y = 1\alpha + x\beta + \epsilon$$

2. GRAMMAR (LD): Aulchenko & al, Genetics, 2007

$$\begin{cases} y = 1\mu_1 + Z_a a + \epsilon_1 \\ \hat{\epsilon_1} = 1\mu_2 + x\beta + \epsilon_2 \end{cases}$$

3. QTDT (LDLA): Abecasis & al, Am. J. Hum. Genet., 2000

$$y = 1\alpha + \frac{x_p + x_m}{2}\beta_b + (x - \frac{x_p + x_m}{2})\beta_w + \epsilon$$

With :
$x = w - \bar{w}$ and $w = \{-2p, (1-2p), 2q\} / \sqrt{2pq}$ for genotype $\{11, 12, 22\}$

Outline

1. Introduction
   - Context
   - Objectives
   - Methods tested
2. Materials & Methods
   - Model and test
   - Model on SNP
3. Results
   - Regression
   - GRAMMAR
   - QTDT
4. Discussion & On going

Regression model and student test (iid)

Regression model :

$$y = 1\alpha + x\beta + \epsilon$$

Student test :

$$\frac{\hat{\beta} - E[\hat{\beta}]}{\sqrt{V(\hat{\beta})}} \sqrt{\frac{E[\hat{\epsilon}'\hat{\epsilon}]}{\hat{\epsilon}'\hat{\epsilon}}} \sim t_{N-2}$$

And in the case of i.i.d we have :

$$\begin{cases} E[\hat{\beta}] = (x'x)^{-} x' E[y] = \beta \\ V(\hat{\beta}) = (x'x)^{-} x' V(y) x (x'x)^{-} = (x'x)^{-} \sigma_{\epsilon}^{2} \\ E[\hat{\epsilon}'\hat{\epsilon}] = (N-2)\sigma_{\epsilon}^{2} \end{cases}$$

$\Rightarrow$ For testing $\beta = 0$ against $\beta \neq 0$, we then use :

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\epsilon}'\hat{\epsilon}}} \sqrt{(N-2)(x'x)} \sim t_{N-2} \overset{N\to\infty}{\longrightarrow} \begin{cases} N(0{,}1)_{H_0} \\ N(\frac{\beta}{\sqrt{V(\hat{\beta})}}{,}1)_{H_1} \end{cases}$$

Regression model and student test (no iid)

Student test :

$$\frac{\hat{\beta} - E[\hat{\beta}]}{\sqrt{V(\hat{\beta})}} \sqrt{\frac{E[\hat{\epsilon}'\hat{\epsilon}]}{\hat{\epsilon}'\hat{\epsilon}}} \sim t_{N-2}$$

And now we have :

$$\left\{\begin{array}{l} E[\hat{\beta}] = (x'x)^{-}x'E[y] = \beta \\ V(\hat{\beta}) = (x'x)^{-}x'V(y)x(x'x)^{-} = (x'x)^{-}\sigma_{\epsilon}^{2} + (x'x)^{-}x'Ax(x'x)^{-}\sigma_{u}^{2} \\ E[\hat{\epsilon}'\hat{\epsilon}] = (N-2)\sigma_{\epsilon}^{2} + (tr(A) - (x'x)^{-}x'Ax - \frac{1}{N}1'A1)\sigma_{u}^{2} \end{array}\right.$$

$\Rightarrow$ For testing $\beta = 0$ against $\beta \neq 0$, we then use :

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\epsilon}'\hat{\epsilon}}} \sqrt{(N-2)(x'x)\frac{(1-h^2) + \frac{h^2}{N-2}(tr(A) - (x'x)^{-}x'Ax - \frac{1}{N}1'A1)}{(1-h^2) + x'Ax(x'x)^{-}h^2}}$$

$$t_{user}\sqrt{\lambda} \overset{N\to\infty}{\longrightarrow} \left\{\begin{array}{l} N(0,1)_{H_0} \\ N(\frac{\beta}{\sqrt{V(\hat{\beta})}},1)_{H_1} \end{array}\right.$$

Regression model and student test (no iid)

So the users use the test :

$$t_{user} \overset{N\rightarrow\infty}{\longrightarrow} \begin{cases} N(0,1/\lambda)_{H_0} \\ N(\frac{\beta}{\sqrt{V(\hat{\beta})\lambda}},1/\lambda)_{H_1} \end{cases}$$

By writing x as $x = w - \bar{w}$, and as $E(w_i w_j) = a_{ij}$ the relationship coefficient between individuals i and j, we found that :

$$E_x[1/\lambda] = \frac{(1-h^2)+C_1 h^2}{(1-h^2)+C_2 h^2}$$

with :

$$\begin{cases} C_1 = (\mu_D - \mu_O) + \frac{N[V_D+(N-1)V_O]}{(N-1)(\mu_D-\mu_O)} \\ \text{and} \\ C_2 = (\mu_D - \mu_O) - \frac{N[V_D+(N-1)V_O]}{(N-2)(N-1)(\mu_D-\mu_O)} \end{cases}$$

D is the diagonal of the relationship matrix A
O is the out diagonal of the relationship matrix A

Outline

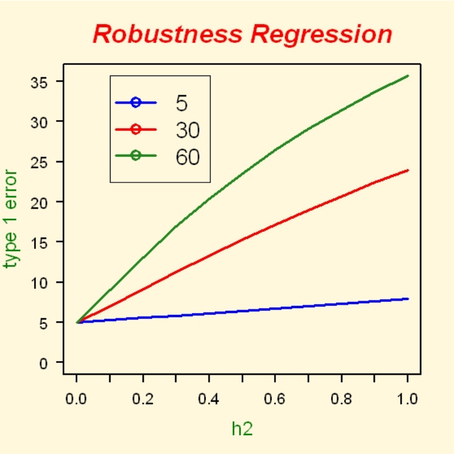
1 Introduction
- Context
- Objectives
- Methods tested

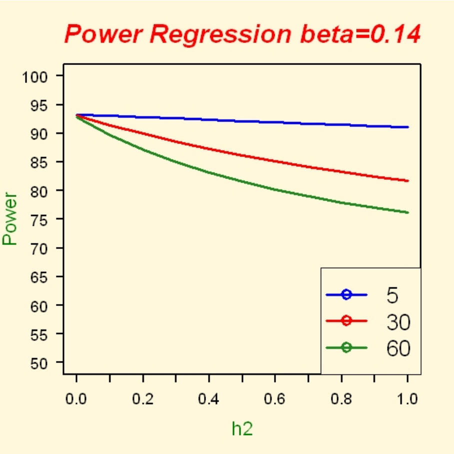
2 Materials & Methods
- Model and test
- Model on SNP

3 Results
- Regression
- GRAMMAR
- QTDT

4 Discussion & On going

Regression

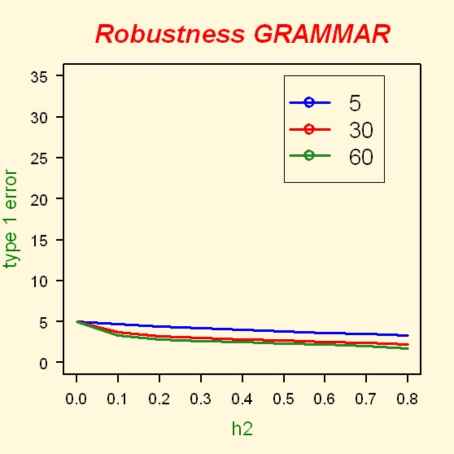


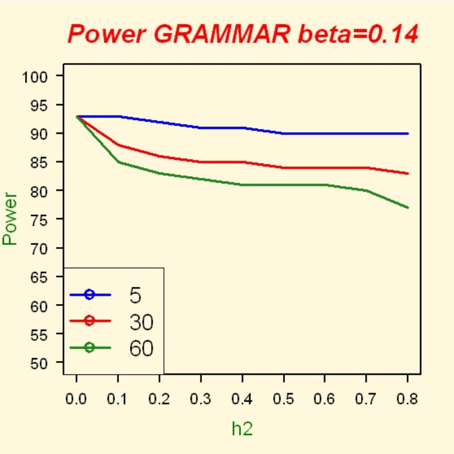


⇒ ↗ structure, ↗ FDR

⇒ ↗ structure, ↘ Power

GRAMMAR

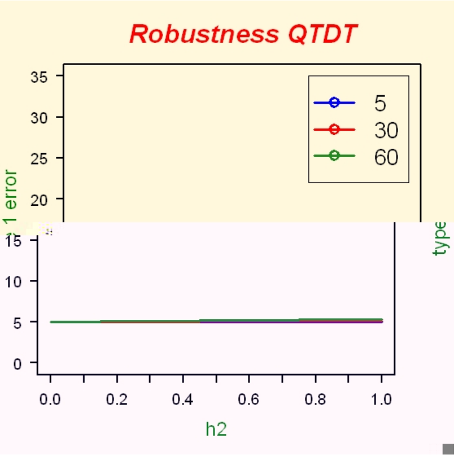


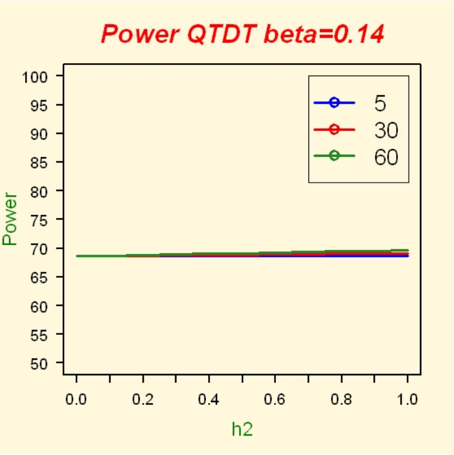


⇒ ↗ structure, ↘ FDR

⇒ ↗ structure, ↘ Power

QTDT





$\Rightarrow$ $\nearrow$ structure, $=$ FDR

$\Rightarrow$ $\nearrow$ structure, $=$ Power

Outline

1. Introduction
   - Context
   - Objectives
   - Methods tested
2. Materials & Methods
   - Model and test
   - Model on SNP
3. Results
   - Regression
   - GRAMMAR
   - QTDT
4. Discussion & On going

Conclusion

- Objective : Provide tools to show that some of these first analysis may give erroneous results in populations with related individuals
- Regression give a lot of problems when there is a parental structure between individuals
- GRAMMAR is more stable but have some problems with the power
- QTDT is stable and power is unaffected, but power isn't high

Acknowledgements

GENEQUIN funders :

Agence Nationale de la Recherche ANR

les Haras nationaux

Région Basse-Normandie

INRA Institut National de la Recherche Agronomique

GENEQUIN partners :

CEMESPO

ECOLE VETERINAIRE ALFORT

INRA Institut National de la Recherche Agronomique

Thank you for your attention !