Incorporating genetic predictors within the SAEM algorithm

Julie Bertrand, Maria de Iorio, David Balding

Genetics Institute, University College London, London, UK
Department of Statistical Science, University College London, London, UK

13 June, 2013

Pharmacogenomics

- Personalized drug therapy[1]
  - High-throughput approach to identifying genetic determinants of drug response
  - lack of large-scale pharmacogenomic studies with adequate follow-up
- *Guideline on the use of pharmacogenetic methodologies in the pharmacokinetic evaluation of medicinal products*[2]
  - large genetic arrays when no hypothesis on genetic origin
  - level of evidence similar to that required in drug-drug interaction
  - modelling and simulation to help in analysis and design

[1] Evans WE, Relling MV. Nature. 2004
[2] EMA/CHMP/37646/2009

Pharmacogenomic model

- Nonlinear mixed effects (NLME)

$$y_{ij} = f(\phi_i, t_{ij}) + \epsilon_{ij} \text{ , with } \epsilon_{ij} \sim N(0, \sigma^2)$$
$$\boldsymbol{\phi_i} = h\left(C_i\boldsymbol{\mu} + \boldsymbol{\eta_i}\right) \text{ , with } \boldsymbol{\eta_i} \sim N(0, \Omega)$$

$h(u) = e^u$ log-normal distribution

$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}, \widehat{\Omega}, \widehat{\sigma})$ $\quad$ $\mathbf{EBE_i} = Argmax_{\phi_i}\ p(\boldsymbol{\phi_i}|\boldsymbol{y_i}; \widehat{\boldsymbol{\theta}})$

Pharmacogenomic model

- Nonlinear mixed effects (NLME)
  - genetic variation: single nucleotide polymorphism, SNP

$$y_{ij} = f(\phi_i, t_{ij}) + \epsilon_{ij} \text{ , with } \epsilon_{ij} \sim N(0, \sigma^2)$$

$$\boldsymbol{\phi_i} = h\left(C_i\boldsymbol{\mu} + \boldsymbol{\eta_i}\right) \text{ , with } \boldsymbol{\eta_i} \sim N(0, \Omega)$$

  - linear regression on allele dosage $SNP = \{0, 1, 2\}$

$$\underset{\boldsymbol{\phi_i}}{\log CL_i} = \underset{C_i}{\begin{pmatrix} 1 & SNP_{1i} & \dots & SNP_{Nsi} \end{pmatrix}} \underset{\boldsymbol{\mu}}{\begin{pmatrix} \mu_{CL} \\ \beta_{CL,SNP_1} \\ \vdots \\ \beta_{CL,SNP_{Ns}} \end{pmatrix}} + \underset{\boldsymbol{\eta_i}}{\eta_{CLi}}$$

$h(u) = e^u$ log-normal distribution

$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}, \widehat{\Omega}, \widehat{\sigma})$ $\quad$ $\mathbf{EBE_i} = Argmax_{\boldsymbol{\phi_i}}\ p(\boldsymbol{\phi_i} | \boldsymbol{y_i}; \widehat{\boldsymbol{\theta}})$

  - number of SNPs, $N_s >> N$, number of subjects
  - varying in informativeness and correlated

Pharmacogenomic analysis

- Method 1: Modified stepwise procedure
  - commonly found in the literature
  - screening step adapted to account for genetic correlation
- Penalised regression
  - established in animal and plant genetics
  - Method 2: Lasso
  - Method 3: HLasso
    - developed for genome-wise association studies
    - higher effect size once included in the model
  - performed on EBE from base model

↪ computationally and statistically efficient[3]

↪ 2-stage approaches: SNP selection after model parameter estimation

[3] Bertrand J, Balding DJ. Pharmacogenet Genomics. 2013

Objectives

- To develop a method 4: integrated approach
  - to simultaneously estimate PK model parameters and genetic effects size
- To compare through a realistic simulation study:
  1. adapted stepwise procedure
  2. Lasso regression on EBE
  3. HLasso regression on EBE
  4. integrated approach

2-stage approaches

1. Stepwise procedure
   i. screening step, for each p^{th} model parameter per SNP
   $\widehat{\beta_{ps}} = argmin_{\beta_{ps}} \sum_i^N (\boldsymbol{EBE_{pi}} - \beta_{ps} \times SNP_{si})^2$
      - pruning on multiple significant SNPs with $r^2 \geq 0.8$
   ii. model inclusion and selection step
   ⟳ repeat i-ii until no more SNPs significant

2-stage approaches

1 Stepwise procedure

- i screening step, for each p^{th} model parameter per SNP
  $\widehat{\beta_{ps}} = argmin_{\beta_{ps}} \sum_i^N (\boldsymbol{EBE_{pi}} - \beta_{ps} \times SNP_{si})^2$
  - pruning on multiple significant SNPs with $r^2 \geq 0.8$
- ii model inclusion and selection step
- ↻ repeat i-ii until no more SNPs significant

- Penalised regression, for each p^{th} model parameter
  $\widehat{\boldsymbol{\beta_p}} = argmin_{\boldsymbol{\beta_p}} \sum_i^N (\boldsymbol{EBE_{pi}} - \boldsymbol{\beta_p} \times \boldsymbol{SNP_i})^2 + P(\boldsymbol{\beta_p})$

2 Lasso, $P_{\xi}(\boldsymbol{\beta_p}) \approx$ double exponential prior on β_p

- ξ set by permutations to ensure a target family wise error rate (FWER)

3 HLasso, $P_{\lambda,\gamma}(\boldsymbol{\beta_p}) \approx$ normal exponential gamma prior on β_p

- λ set to 1, γ set by permutations

Integrated approach

- Simultaneous SNP selection and estimation of PK model parameters
  - HLasso at each iteration of the SAEM algorithm
- Maximization-step of $\boldsymbol{\mu}$ in SAEM

$\widehat{\boldsymbol{\mu}_{\boldsymbol{k+1}}} = argmin_{\boldsymbol{\mu}} \sum_{i=1}^{N} (\boldsymbol{s_{ik}} - C_i\boldsymbol{\mu})'\Omega^{-1}(\boldsymbol{s_{ik}} - C_i\boldsymbol{\mu})$

At iteration k

$\boldsymbol{\phi_{ik}}$ drawn from $p(.|\mathbf{y}; \boldsymbol{\theta_k})$

$\boldsymbol{s_{ik}} = \boldsymbol{s_{ik-1}} + \tau_k(\boldsymbol{\phi_{ik}} - \boldsymbol{s_{ik-1}})$

$\mu = (\mu_{Cl}, \mu_V, \beta_{Cl,1}, \ldots, \beta_{Cl,N_s})$

τ_k, a decreasing sequence of positive numbers

Integrated approach

- Simultaneous SNP selection and estimation of PK model parameters
  - HLasso at each iteration of the SAEM algorithm
- Maximization-step of $\boldsymbol{\mu}$ in the integrated approach

$$\widehat{\boldsymbol{\mu}_{\boldsymbol{k+1}}} = argmin_{\boldsymbol{\mu}} \sum_{i=1}^{N} (\boldsymbol{s_{ik}} - C_i\boldsymbol{\mu})'\Omega^{-1}(\boldsymbol{s_{ik}} - C_i\boldsymbol{\mu}) + P_{\lambda,\gamma}(\boldsymbol{\mu})$$

  - call to `hlasso` program with $\boldsymbol{s_{ik}}$ as the response
  - λ set to 1, γ set using an asymptotic approximation
  - implemented in the `saemix` R package

At iteration k

ϕ_{ik} drawn from $p(.|\mathbf{y};\boldsymbol{\theta_k})$

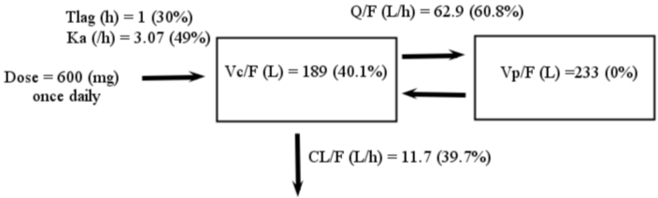
$\boldsymbol{s_{ik}} = \boldsymbol{s_{ik-1}} + \tau_k(\phi_{ik} - \boldsymbol{s_{ik-1}})$

$\mu = (\mu_{Cl}, \mu_V, \beta_{Cl,1}, \dots, \beta_{Cl,N_s})$

τ_k, a decreasing sequence of positive numbers

Pharmacokinetic settings

- Structural and statistical model
  - inspired from real study [4]

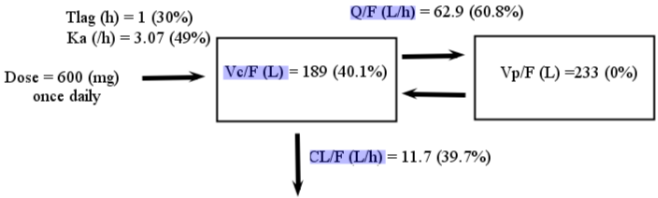


  - diagonal variance matrix of random effects
  - combined residual error model
- Phase II-like study design
  - 300 individuals with t= 0.5, 1.25, 2, 4, 9, 24

[4] Kappelhoff et al. Clinical pharmacokinetics, 2005

Pharmacokinetic settings

- Structural and statistical model
  - inspired from real study [4]



  - diagonal variance matrix of random effects
  - combined residual error model
- Phase II-like study design
  - 300 individuals with t= 0.5, 1.25, 2, 4, 9, 24

[4] Kappelhoff et al. Clinical pharmacokinetics, 2005

Genetic settings

- Generation of genotypes using HAPGEN [5]
  - N_s=1227 snps on 171 genes from the DMET Chip [6]
  - 6 [1-56] snps per gene
  - HAPMAP caucasian reference haplotypes
- Alternative hypothesis H_1=presence of a genetic effect
  - 200 simulated data sets
  - 6 *unobserved* causal variants with allele frequency, p_s
    - decrease in log(CL/F) with allele dosage
    - varying genetic component of interindividual variability

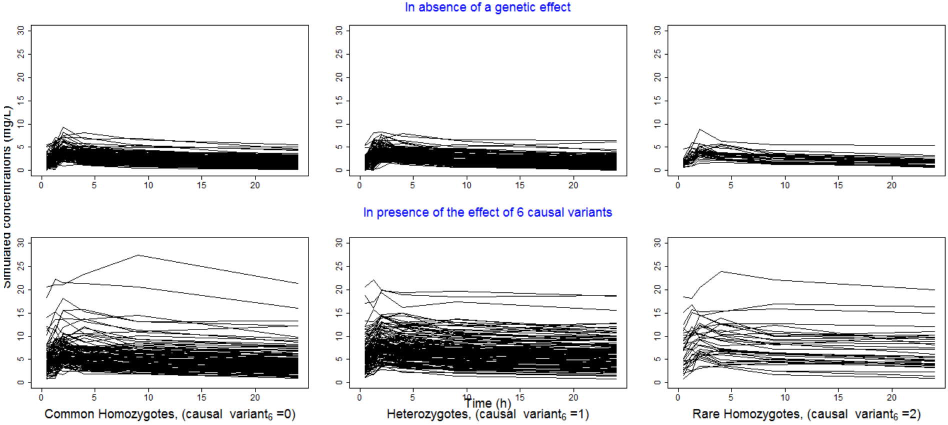
$$R_{Gs} = \frac{\beta_s^2 \times 2p_s(1-p_s)}{\beta_s^2 \times 2p_s(1-p_s) + \omega_{CL/F}^2} = (1,2,3,5,7,12)'\,\%$$

$$R_G = \sum_{s=1}^{6} R_{Gs} = 30\%$$

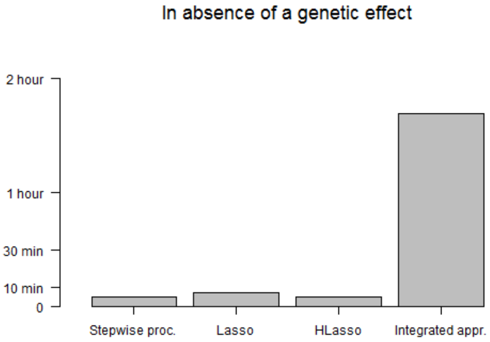
[5] Su et al. Bioinformatics, 2011

[6] Daly et al. Clinical Chemistry, 2007

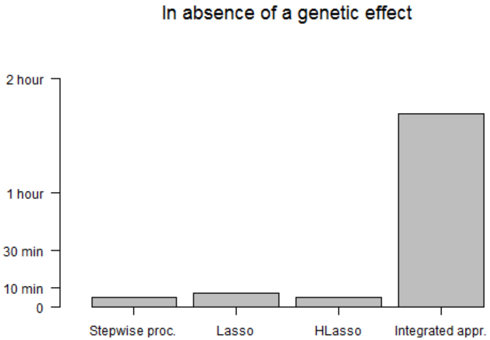
A typical simulated dataset



Computing times

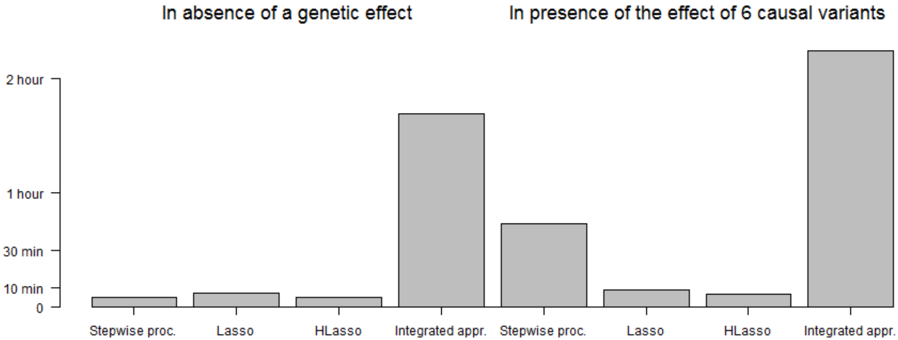


Computing times



- Similar computing times for 2-stage approaches
- Integrated approach
  - HLasso run at each SAEM iteration

Computing times



- Similar computing times for 2-stage approaches
- Integrated approach
  - HLasso run at each SAEM iteration
- Slight increase for all methods
- Stepwise proc.= 10 times longer run times under H_1

FWER and TP

	FWER(%)	TP	$FP_{CL/F}$	$FP_{Vc/F}$	$FP_{Q/F}$
Stepwise proc.	18.5	338 [302–374]	15 [7–23]	8 [2–14]	30 [19–41]
Lasso	18.5	311 [276–346]	12 [5–19]	18 [10–26]	11 [4–18]
HLasso	18	316 [281–351]	14 [7–21]	15 [7–23]	11 [4–18]
Integrated appr.	20	256 [225-287]	19 [10-28]	7 [2-12]	0

Family wise error rate, FWER= expected value of 20[14.5–25.5]%

True positive, TP=SNP in $r^2 \geq 0.05$ with causal variant; maximum, possible 1200

False positives, FP

FWER and TP

	FWER(%)	TP	$FP_{CL/F}$	$FP_{Vc/F}$	$FP_{Q/F}$
Stepwise proc.	18.5	338 [302–374]	15 [7–23]	8 [2–14]	30 [19–41]
Lasso	18.5	311 [276–346]	12 [5–19]	18 [10–26]	11 [4–18]
HLasso	18	316 [281–351]	14 [7–21]	15 [7–23]	11 [4–18]
Integrated appr.	20	256 [225-287]	19 [10-28]	7 [2-12]	0

Family wise error rate, FWER= expected value of 20[14.5–25.5]%

True positive, TP=SNP in $r^2 \geq 0.05$ with causal variant; maximum, possible 1200

False positives, FP

- target FWER of 20% achieved with all methods

FWER and TP

	FWER(%)	TP	$FP_{CL/F}$	$FP_{Vc/F}$	$FP_{Q/F}$
Stepwise proc.	18.5	338 [302–374]	15 [7–23]	8 [2–14]	30 [19–41]
Lasso	18.5	311 [276–346]	12 [5–19]	18 [10–26]	11 [4–18]
HLasso	18	316 [281–351]	14 [7–21]	15 [7–23]	11 [4–18]
Integrated appr.	20	256 [225-287]	19 [10-28]	7 [2-12]	0

Family wise error rate, FWER= expected value of 20[14.5–25.5]%

True positive, TP=SNP in $r^2 \geq 0.05$ with causal variant; maximum, possible 1200

False positives, FP

- target FWER of 20% achieved with all methods
- Integrated approach
  - lower TP count

FWER and TP

	FWER(%)	TP	$FP_{CL/F}$	$FP_{Vc/F}$	$FP_{Q/F}$
Stepwise proc.	18.5	338 [302–374]	15 [7–23]	8 [2–14]	30 [19–41]
Lasso	18.5	311 [276–346]	12 [5–19]	18 [10–26]	11 [4–18]
HLasso	18	316 [281–351]	14 [7–21]	15 [7–23]	11 [4–18]
Integrated appr.	20	256 [225-287]	19 [10-28]	7 [2-12]	0

Family wise error rate, FWER= expected value of 20[14.5–25.5]%

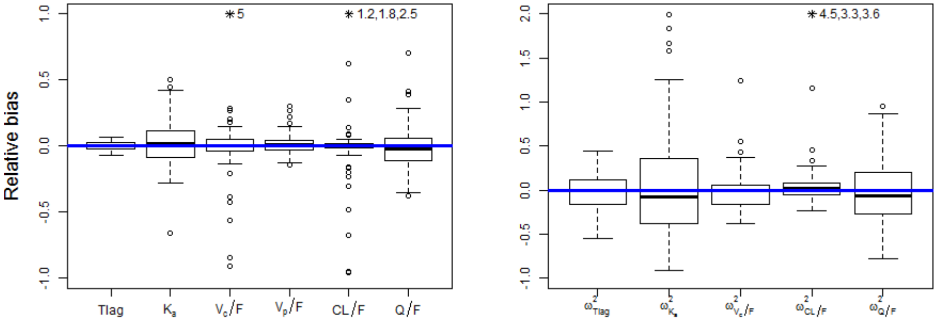
True positive, TP=SNP in $r^2 \geq 0.05$ with causal variant; maximum, possible 1200

False positives, FP

- target FWER of 20% achieved with all methods
- Integrated approach
  - lower TP count
  - lower FP count on Vc/F and Q/F

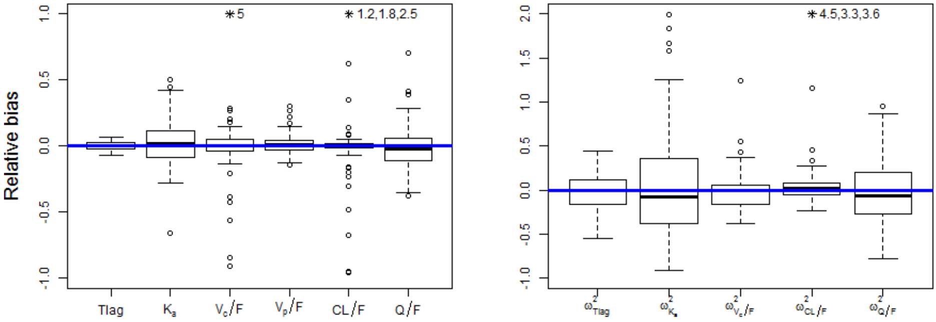
Estimation performance

Integrated approach in absence of a genetic effect



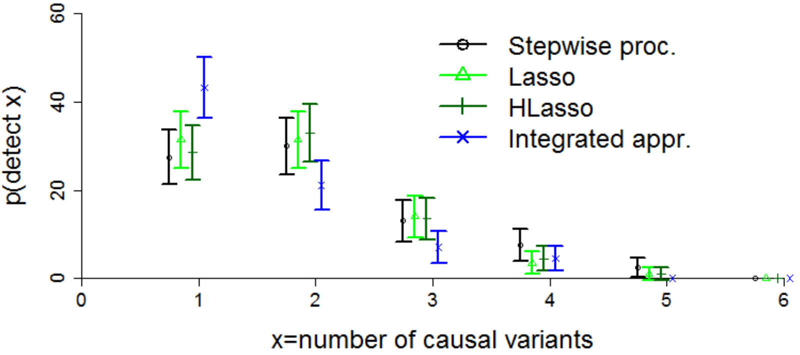
Estimation performance

Integrated approach in absence of a genetic effect

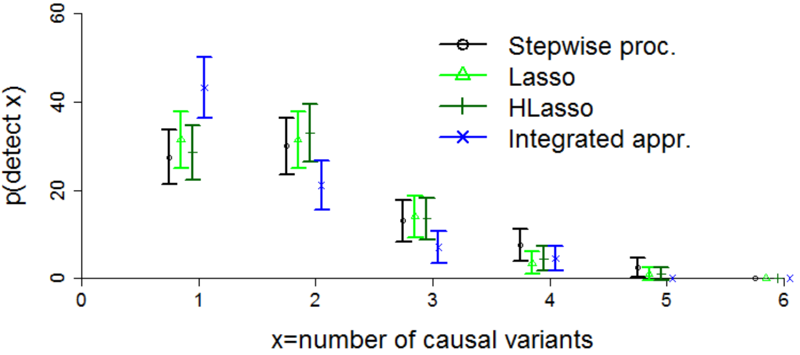


- Fixed effects
  - less than 3% Rbias and RRMSE from 3-15%
- Variances
  - less than 5% Rbias and RRMSE from 20-50%

Power to detect multiple variants

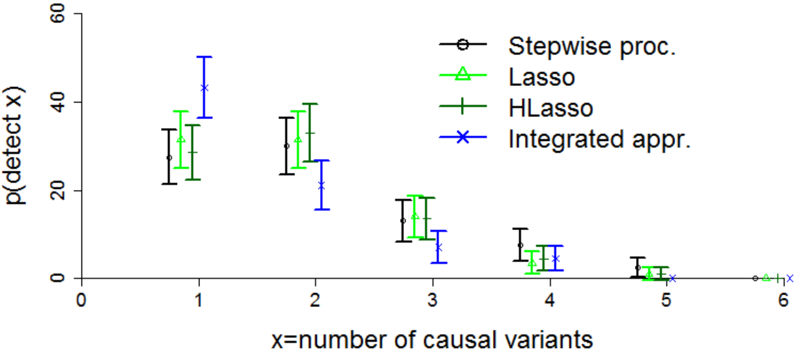


Power to detect multiple variants



- None of the approaches select the 6 causal variants

Power to detect multiple variants



- None of the approaches select the 6 causal variants
- Integrated approach favours more parsimonious models

Discussion

- Realistic simulation study
  - feasability of combining large SNPs set and NLME model
  - chosen FWER of 20% to enable power comparisons
  - analyses for exploratory purposes
    - further functional studies required
- Integrated approach
  - \+ full model-based approach
  - \+ less false positives
  - – longer computing times
  - – less powerful to detect multiple SNPs
- Future works
  - influence of shape parameter
    - larger shape parameter → Lasso
  - full Bayesian approach

Acknowledgements

- `saemix` R package
  Dr Emmanuelle Comets
- UCL Genetics Institute
- London Pharmacometrics Interest Group

Asymptotic approximation to set γ

$$\frac{sign(\beta_p = 0^+)(2\lambda + 1)}{\gamma} \frac{D_{-(2\lambda+2)}\left(\frac{|\beta_p = 0^+|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta_p = 0^+|}{\gamma}\right)} = \Phi^{-1}(1 - \alpha/2)\sqrt{\frac{N}{\delta_p}}$$

$\delta_p = \mathrm{VAR}(s_{p.k})/\omega_p^2$

reflects the design information

$\mathrm{VAR}(s_{p.k}) \ll \omega_p^2 \rightarrow$ increases penalisation

$\mathrm{VAR}(s_{p.k})$ derived using Batch means method