Identification and characterization of individual variability within a population based on a mechanistic model and mixed effects. Application in breeding

Estelle Kuhn

Applied Mathematics and Computer Science from Genome to Environment
Jouy-En-Josas

anr Stat4Plant

Joint works mainly with C. Baey, M.Delattre, J.B. Leger, S. Lemler, T. Guédon, A. Caillebotte

Agriculture's transition

⇒ climate change, limited resource, demographic evolution, economic constraints, ...

New objectives for agriculture

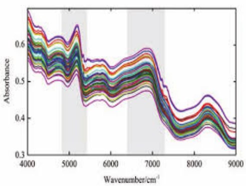
- multivariate traits performance → global system analysis
- robustness, resilience → global/local adaptation properties

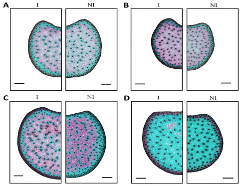
⇒ require finely understanding underlying processes

- new available technologies, new data





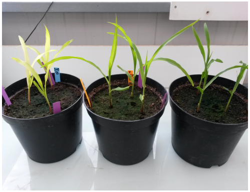




Plant breeding

Genotype by Environment effect



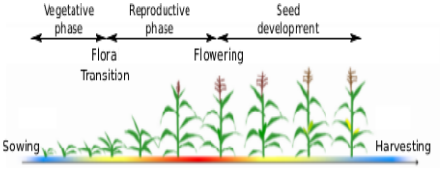


⇒ Strong interaction between genotype and environment (climat, soil, crop managment, ...)

Challenges :

- capitalize on genotype by environment interactions to find well-adapted genotypes
- target multi-objective performance
- integrate biological knowledge through modeling
- ...

Plant growth process

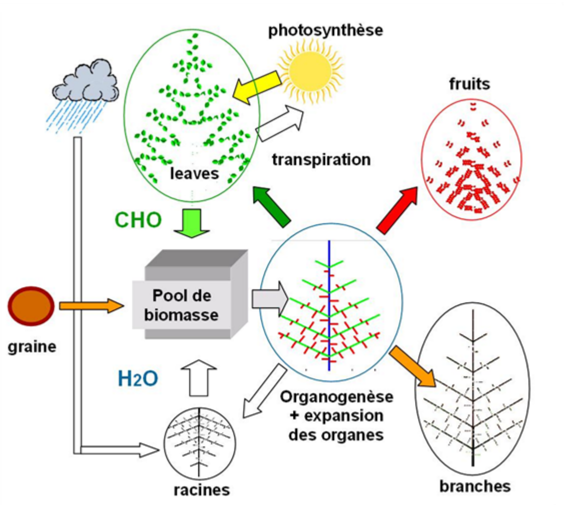


- Integrated and longitudinal quantities of interest, times of interest
- Numerous covariates (temperature, rainfall, soil composition)

$\Longrightarrow$ description of processes

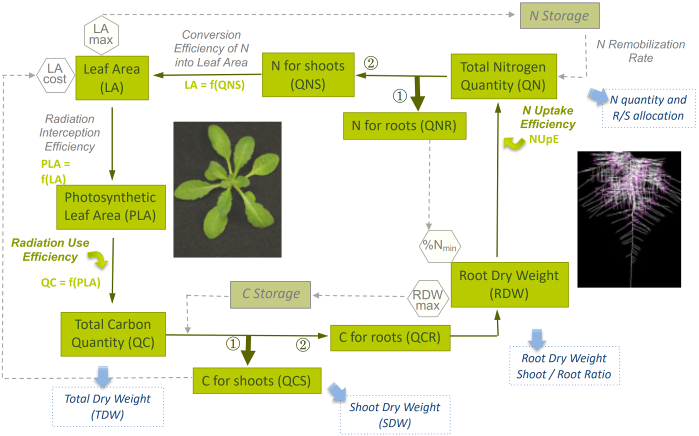
$\Longrightarrow$ plant ecophysiology

Crop growth modeling



Arnica model

[Richard-Molard et al. (2007)]
⇒ modeling carbon and nitrogen flow in *Arabidopsis Thaliana*



Observations of growing process along time

[Pinheiro and Bates (2000)]

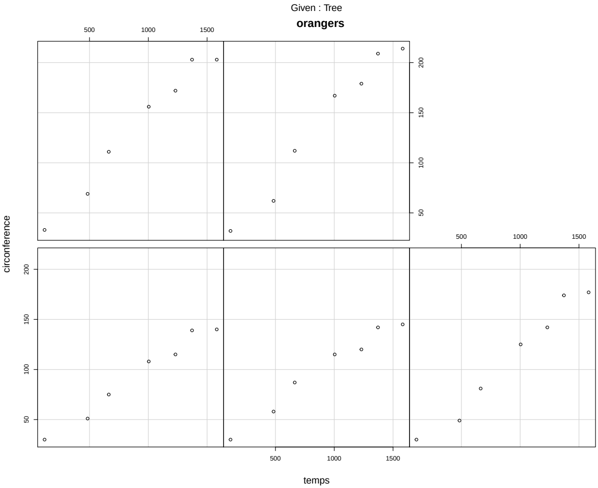
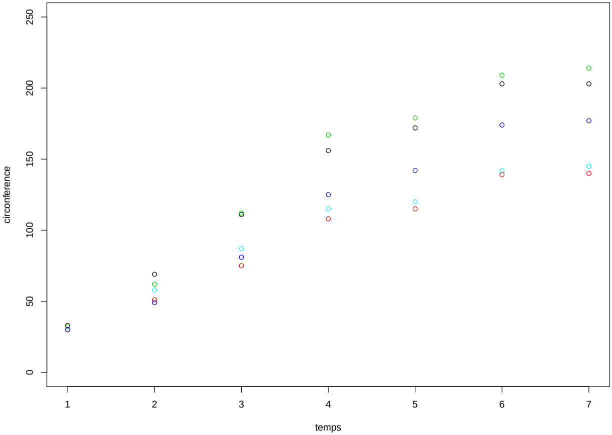


Figure: Circumferences of 5 orange trees measured at 7 times

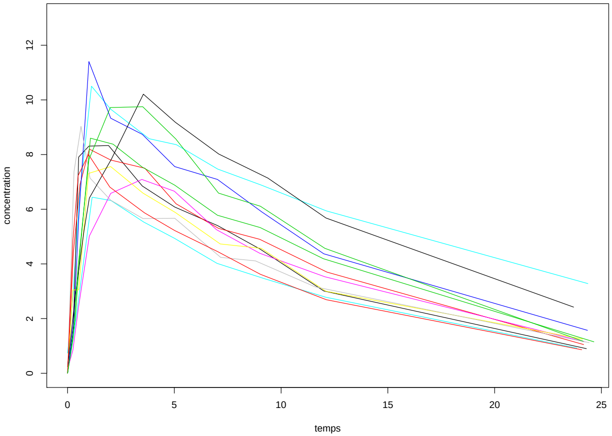
Observations of growing process of five orange trees



logistic model $y(t) = \frac{\varphi_{i1}}{1+\exp\left(-\frac{t-\varphi_{i2}}{\varphi_{i3}}\right)}$

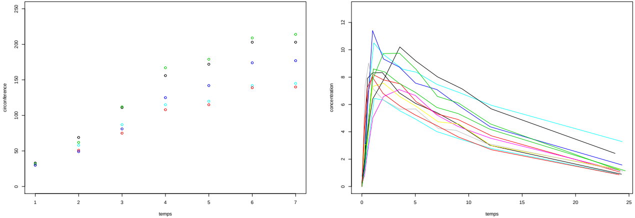
Theophylline concentration along time

[Davidian and Giltinan (1995)]



12 subjects, same oral dose (mg/kg) times in hours theophylline concentration in mg/L

Biological objectives



- Understanding intra-subject processes
- Understanding variations of these processes across subjects

$\Rightarrow$ fundamental for developing individual strategies and guidelines

General context

- Repeated measurements over time (or other conditions) within individuals from a population of interest
- A model for individual profiles with interpretable parameters available
- Inference focuses on mechanisms that underlie individual profiles and variations in the population

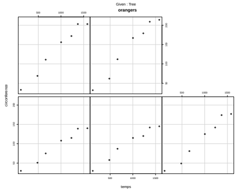
Let Y_{ij} be the observation at the jth measurement for individual i for $1 \leq j \leq J$ and $1 \leq i \leq N$
Example of orange trees

$$Y_{ij} = \frac{\varphi_{i1}}{1 + \exp\left(-\frac{t_j - \varphi_{i2}}{\varphi_{i3}}\right)} + \varepsilon_{ij}$$

where Y_{ij} is circumference of tree i at time t_j and φ_i parameters of tree i

Individual level approach versus population approach

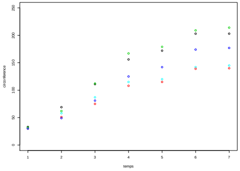
- Ajusting N regression models each with J observations



$$Y_{ij} = h(\varphi_i, t_j) + \varepsilon_{ij}$$

Model parameters : $(\varphi_i, \sigma_i^2) \in \mathbb{R}^{d+1}$
$\Rightarrow N(d+1)$ parameters

- Ajusting 1 model with NJ observations



$$\begin{cases} Y_{ij} & = & h(\varphi_i, t_j) + \varepsilon_{ij} \\ \varphi_i & \sim & \mathcal{L}(\nu) \end{cases}$$

Model parameters : $\theta_{pop} = (\nu, \sigma^2)$

Mixed effect model: art of modeling variabilities ?

- modeling observations conditionaly to individual parameter
  $\Longrightarrow$ individual level model

$$y_{ij} = h(\alpha, \varphi_i, t_j) + \varepsilon_{ij}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq J,$$

  with y_{ij} measurement of individual i in environment j
  φ_i parameter of individual i
  t_j environnemental covariates
  α population parameters vector
  Σ noise parameter
- modeling variability of model parameter using individual parameter
  $\Longrightarrow$ population level model

$$\varphi_i = \beta + b_i \text{ with } b_i \sim \mathcal{N}(0; \Gamma), \quad 1 \leq i \leq N,$$

Statistical issues raised up

Consider the following mixed effects model:

$$\begin{cases} Y_{ij} = h(\alpha, \varphi_i, t_j) + \varepsilon_{ij} & 1 \leq i \leq N,\ 1 \leq j \leq J \\ \varphi_i = \beta + b_i, & 1 \leq i \leq N \end{cases}$$

with $b_i \overset{iid}{\sim} \mathcal{N}(0; \Gamma)$ random effects, $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0; \Sigma)$ noise term

Objectives:

- estimate model parameters $\theta = (\alpha, \beta, \Gamma, \Sigma)$ $\Rightarrow$ Focus 1
- predict individual output as $\hat{\varphi}_i$ or $\hat{Y}_i$
- test if some random effects (b_i) are null $\Rightarrow$ Focus 2
- explain variabilities of individual parameters φ_i $\Rightarrow$ Focus 3
- ...

Focus 1: Inference in mixed effects model

Consider the following mixed effects model:

$$\begin{cases} Y_{ij} &=& h(\alpha, \varphi_i, t_j) + \varepsilon_{ij} \quad 1 \le i \le N,\ 1 \le j \le J \\ \varphi_i &=& \beta + b_i, \quad\quad\quad\quad\quad 1 \le i \le N \end{cases}$$

with $b_i \overset{iid}{\sim} \mathcal{N}(0; \Gamma)$ random effects, $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0; \Sigma)$ noise term
Model parameter $\theta = (\alpha, \beta, \Gamma, \Sigma) \in \Theta$

Complete likelihood of individual i:

$$L_{comp}(\theta; Y_i, b_i) \;=\; f(Y_i | b_i; \alpha, \beta, \Sigma) f(b_i; \Gamma)$$

$\Rightarrow$ the random effects (b_i) are non observed

Observed marginal likelihood:

$$L_{marg}(\theta; Y_i) \;=\; \int L_{comp}(\theta; Y_i, b_i) db_i$$

Define the maximum likelihood estimate (MLE) by:

$$\hat{\theta}_N = \arg\max_{\theta \in \Theta} L_{marg}(\theta; Y_1^N)$$

Compute the maximum likelihood estimate

$$\widehat{\theta} = \arg\max_{\theta} L_{marg}(y;\theta) = \arg\max_{\theta} \int L_{comp}(y,z;\theta)\, dz$$

Main tools:

- Expectation Maximization like algorithms:
  - EM algorithm (Dempster et al. (1977); Wu (1983); Balakrishnan et al. (2017))
  - stochastic versions of EM: Stochastic EM (Celeux et al 1995), Monte Carlo EM, (Fort et al 2003), stochastic approximation EM, (Delyon et al. (1999); Allassonnière et al. (2007))
  - variational versions of EM (Bernardo et al 2003)

  Main limitations:
  - theoretical results in exponential family
  - computationaly tricky out of exponential family
  - target can be different from MLE using variational EM or exponentialization trick (Debavelaere and Allassonnière (2021))
- gradient based method
  $\rightarrow$ stochastic gradient like algorithm (Cappé et al. (2005))

Stochastic gradient algorithm

Objective: compute $\widehat{\theta} = \arg\max_\theta L_{marg}(y;\theta)$

Gradient algorithm : maximizing $g(\theta)$ iteratively through

$$\theta_{k+1} = \theta_k + \gamma_k \nabla_\theta g(\theta_k)$$

Stochastic gradient algorithm:

If $\hat{\nabla_\theta} g(\theta, Z)$ is an estimate of $\nabla_\theta g(\theta)$ maximizing $g(\theta)$

for $k = 1, \cdots$ **do**
  $z_k \leftarrow$ random sample from Z
  $\theta_{k+1} = \theta_k + \gamma_k \hat{\nabla_\theta} g(\theta_k, Z_k)$
end for

Fisher identity in latent variable model

Observed log-likelihood: $\log g(y;\theta) = \log \int f(y,z;\theta)dz$

Fisher identity:

$$\nabla_\theta \log g(y;\theta) = \mathrm{E}(\nabla_\theta \log f(y,Z;\theta) \mid y;\theta)$$

$\Rightarrow$ compute $\widehat{\theta} = \arg\max_\theta \log g(y;\theta)$ iteratively

for $k = 1, \cdots$ **do**

$z_k \leftarrow$ random sample from $p(. \mid y;\theta_k)$

$\theta_{k+1} = \theta_k + \gamma_k \nabla_\theta \log f(y, z_k;\theta_k) \mid y;\theta_k)$

end for

Preconditioning by Fisher information matrix

$$\mathcal{I}(\theta) = \mathrm{E}\left[(\nabla_\theta \log g(Y;\theta))(\nabla_\theta \log g(Y;\theta))^T\right]$$

- preconditionning the gradient allow a large speed-up
- caracterizing the MLE

$\Rightarrow$ estimate $\mathcal{I}(\theta)$ for $(y_1, \ldots, y_n)$ independent with

$$\widehat{\mathcal{I}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log g(y_i;\theta)(\nabla_\theta \log g(y_i;\theta))^T$$

following Delattre and Kuhn (2023) and using again Fisher identity

$$\widehat{\mathcal{I}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\nabla_\theta \log f(y_i, z_i;\theta) \mid y_i;\theta\right] \mathrm{E}\left[\nabla_\theta \log f(y_i, z_i;\theta) \mid y_i;\theta\right]^T$$

Advantages: $\widehat{\mathcal{I}}(\theta) \geq 0$, no additional cost

The algorithm Fisher-SGD

[Baey et al. (2023)]

for $k = 1, \ldots, K$ **do**
    for $i = 1, \ldots, N$ **do**
        $z_i^k \leftarrow$ sample from $p_{\theta_{k-1}}(\cdot \mid y_i)$
    end for

    $v_k \leftarrow \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \log f\left(y_i, z_i^k; \theta_k\right)$

    for $i = 1, \ldots, N$ **do**
        $\Delta_i^k \leftarrow (1 - \gamma_k)\Delta_i^{k-1} + \gamma_k \nabla_\theta \log f\left(y_i, z_i^k; \theta_k\right)$
    end for
    $I_k \leftarrow \frac{1}{N} \sum_{i=1}^{N} \Delta_i^k \left(\Delta_i^k\right)^T$
    $\theta_{k+1} \leftarrow \theta_k + \gamma_k I_k^{-1} v_k$
end for
Output: θ_k, I_k

Theoretical result

Let $F(\theta) = -\log g(y;\theta)$

Theorem: Under regularity assumptions, and assuming Θ bounded, the iterates $(\theta_k)_k$ defined in Fisher-SGD satisfy

$$\mathrm{E}\left[\min_{0\leq l\leq k} \|\nabla_\theta F(\theta_l)\|^2\right] \leq \square \frac{(F(\theta_0) - \min F)}{\sum_{l=0}^k \gamma_l} + \square \frac{\sum_{l=0}^k \gamma_l^2}{\sum_{l=0}^k \gamma_l}.$$

Application to nonlinear mixed effect model

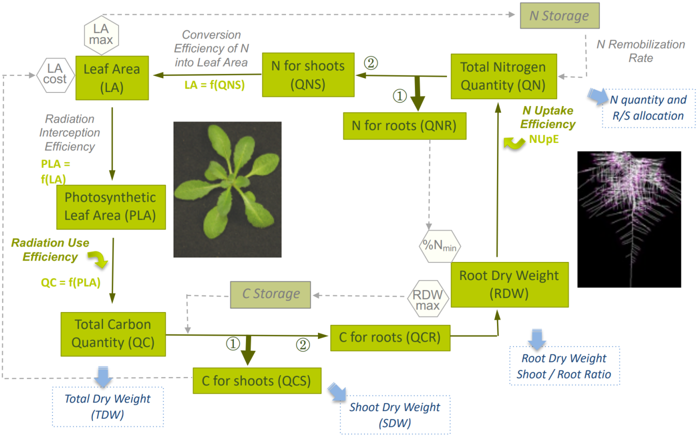
$$\begin{cases} Y_{ij} \mid \varphi_i & \sim & \mathcal{N}\left(\frac{Z_{i1}}{1+\exp\left(-\frac{t_{ij}-\varphi_{i2}}{\alpha}\right)}, \sigma^2\right) \\ \varphi_i & \sim & \mathcal{N}(\beta, \Gamma) \end{cases}$$

Parameters: $\theta = \left(\alpha, \beta, \Gamma, \sigma^2\right)$
comparison with MCMC-SAEM which use exponentialisation trick and block-diagonal FIM estimate

Type	Fisher-SGD		MCMC-SAEM	
	RMSE	Coverage	RMSE	Coverage
β_1	0.234	0.942 ± 0.012	0.236	0.941 ± 0.015
β_2	0.586	0.958 ± 0.010	0.625	0.941 ± 0.015
α	0.414	0.972 ± 0.013	0.416	0.968 ± 0.011
Γ_{11}	2.221	0.951 ± 0.013	2.241	0.949 ± 0.014
Γ_{12}	4.156	0.948 ± 0.014	4.334	0.935 ± 0.015
Γ_{22}	14.324	0.948 ± 0.014	16.492	0.905 ± 0.018
σ^2	1.005	0.957 ± 0.012	1.010	0.951 ± 0.013

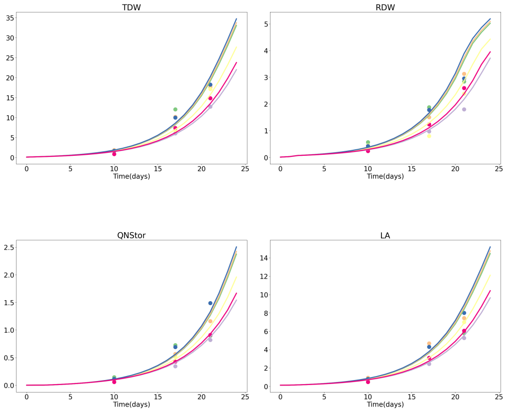
Arnica model

[Richard-Molard et al. (2007)]
⇒ modeling carbon and nitrogen flow in *Arabidopsis Thaliana*



Individual prediction

[Tom Guédon's PhD]



Take home message

- Fisher-SGD performs parameter estimation
  in general latent variable models.
  $\Rightarrow$ Tom Rohmer's talk
- efficient preconditioning through Fisher information matrix.
- simultaneously estimate FIM for free
- easy to implement and generic tuning rules are provided.
- theoretical guarantees in a wide range of latent variable models.

Baey, C., Delattre, M., Kuhn, E., Leger, J.B., Lemler, S. (2023). Efficient preconditioned stochastic gradient descent for estimation in latent variable models. *International Conference on Machine Learning*

Focus 2: Identifying individual variabilities among the population

$$\begin{cases} Y_{ij} = h(\alpha, \varphi_i, t_j) + \varepsilon_{ij} & 1 \leq i \leq N,\ 1 \leq j \leq J \\ \varphi_i = \beta + b_i, & 1 \leq i \leq N \end{cases}$$

with $b_i \overset{iid}{\sim} \mathcal{N}(0; \Gamma)$ random effects, $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0; \Sigma)$ noise term

$\Rightarrow$ *Test for variance components in mixed effects model*

Objective: test that r random effects among p have null variances.

Let $\Gamma = \left(\begin{array}{c|c} \Gamma_1 & \Gamma_{12} \\ \hline \Gamma_{12}^t & \Gamma_2 \end{array}\right)$ where $\Gamma_1 \in \mathcal{S}_{p-r}^+$ and $\Gamma_2 \in \mathcal{S}_r^+$

$$\Theta_0 = \{\theta \in \mathbb{R}^q | \beta \in \mathbb{R}^p, \Gamma_1 \in \mathcal{S}_{p-r}^+, \Gamma_2 = 0, \Gamma_{12} = 0, \Sigma \in \mathcal{S}_J^+\}$$
$$\Theta_1 = \{\theta \in \mathbb{R}^q | \beta \in \mathbb{R}^p, \Gamma \in \mathcal{S}_p^+, \Sigma \in \mathcal{S}_J^+\}$$

$\Longrightarrow$ test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$

Asymptotic distribution of the LRT statistic

The likelihood ratio test statistic equals to

$$LRT_N = -2\log\left(\frac{\sup_{\theta\in\Theta_0} L_N(\theta)}{\sup_{\theta\in\Theta_1} L_N(\theta)}\right) = 2(\ell_N(\hat{\theta}_{H_1}) - \ell_N(\hat{\theta}_{H_0}))$$

with $L_N(\theta) = \prod_1^N f_\theta(Y_i)$ for $(Y_1, ..., Y_N)$ a sample.

Consider the test defined by H_0 : "$R\theta = 0$" against H_1 : "$R\theta \neq 0$"

where R is a full rank matrix of size rxp.

Then, assuming regularity conditions, under H_0:

$$LRT_N = -2\log\left(\frac{\sup_{\theta\in\Theta_0} L_N(\theta)}{\sup_{\theta\in\Theta_1} L_N(\theta)}\right) = 2(\ell_N(\hat{\theta}_{H_1}) - \ell_N(\hat{\theta}_{H_0})) \xrightarrow{\mathcal{L}} \chi^2(r)$$

Asymptotic distribution of the LRT statistic for linear hypotheses defined by inequalities when Θ is open

[Self and Liang (1987)]
Consider the test defined by H_0 : "$R\theta = 0$" against H_1 : "$R\theta \geq 0$"
where R is a full rank matrix

Denote by I_0 the corresponding Fisher information matrix.

Then, assuming regularity conditions, under H_0:

$$LRT_n \xrightarrow{\mathcal{L}} \min_{R\theta=0} (Z-\theta)^t I_0 (Z-\theta) - \min_{R\theta\geq 0} (Z-\theta)^t I_0 (Z-\theta)$$

where $Z \sim \mathcal{N}(0, I_0^{-1})$

Limits of existing results

Example of testing one variance equals to zero considering two correlated random effects:

Let $\theta = (\beta, \Gamma, \Sigma)$ with $\Gamma = \begin{pmatrix} \gamma_1^2 & \gamma_{12} \\ \gamma_{12} & \gamma_2^2 \end{pmatrix}$ and $\Theta = \mathbb{R}^2 \times \mathcal{S}_2^+ \times \mathcal{S}_J^+$.

Consider $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ with

$$\Theta_0 = \{\theta, \beta \in \mathbb{R}^2, \gamma_1^2 = \gamma_{12} = 0, \gamma_2^2 \geq 0, \Sigma \in \mathcal{S}_J^+\}$$
$$\Theta_1 = \{\theta, \beta \in \mathbb{R}^2, \gamma_1^2 \geq 0, \gamma_1^2\gamma_2^2 - \gamma_{12}^2 \geq 0, \gamma_2^2 \geq 0, \Sigma \in \mathcal{S}_J^+\}$$

$\Longrightarrow$ Θ is not open

$\Longrightarrow$ general hypotheses

Identifying the asymptotic distribution of the LRT statistics for testing variance components in nonlinear mixed effects model

[Baey et al. (2019)]

Consider the test defined by
$H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ where
$\Theta_0 = \{\theta \in \mathbb{R}^q | \beta \in \mathbb{R}^p, \Gamma_1 \in \mathcal{S}^+_{p-r}, \Gamma_2 = 0, \Gamma_{12} = 0, \Sigma \in \mathcal{S}^+_J\}$
$\Theta_1 = \{\theta \in \mathbb{R}^q | \beta \in \mathbb{R}^p, \Gamma \in \mathcal{S}^+_p, \Sigma \in \mathcal{S}^+_J\}$
Then, assuming regularity assumptions, under H_0:

$$LRT_n \xrightarrow{\mathcal{L}} \bar{\chi}^2(I_0^{-1}, T(\Theta_0, \theta_0)^\perp \cap T(\Theta_1, \theta_0)),$$

where $T(\Theta, \theta)$ is the tangent cone of Θ at θ and $\bar{\chi}^2(V, \mathcal{C})$ has a χ-bar square distribution (mixture of chi square distributions) with $\mathcal{C}$ a closed convex cone and V a positive definite matrix

Example of testing one variance equals to zero considering two random effects

Let $\theta = (\beta, \Gamma, \Sigma)$

- independent case: $\Gamma = \begin{pmatrix} \gamma_1^2 & 0 \\ 0 & \gamma_2^2 \end{pmatrix}$

  Consider $H_0 : \gamma_1^2 = 0$ against $H_1 : \gamma_1^2 \geq 0$

$$LRT_n \xrightarrow{d} \frac{1}{2}\chi^2(0) + \frac{1}{2}\chi^2(1)$$

- correlated case: $\Gamma = \begin{pmatrix} \gamma_1^2 & \gamma_{12} \\ \gamma_{12} & \gamma_2^2 \end{pmatrix}$

  Consider $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$

$$\Theta_0 = \{\theta, \beta \in \mathbb{R}^2, \gamma_1^2 = \gamma_{12} = 0, \gamma_2^2 \geq 0, \Sigma \in \mathcal{S}_J^+\}$$
$$\Theta_1 = \{\theta, \beta \in \mathbb{R}^2, \gamma_1^2 \geq 0, \gamma_1^2\gamma_2^2 - \gamma_{12}^2 \geq 0, \gamma_2^2 \geq 0, \Sigma \in \mathcal{S}_J^+\}$$

$$LRT_n \xrightarrow{d} \frac{1}{2}\chi^2(1) + \frac{1}{2}\chi^2(2)$$

Empirical level of the test for one effect when two effects are correlated in the linear model

$$Y_{ij} = \varphi_{1i} + \varphi_{2i} t_{ij} + \varepsilon_{ij} \ ,$$

Let $\Gamma = \begin{pmatrix} \gamma_1^2 & \gamma_{12} \\ \gamma_{12} & \gamma_2^2 \end{pmatrix}$
Consider $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$

Table: Percentages of rejection for the LRT procedure for $n = 500$ for the nominal level of the test α on 300 repetitions.

α	$\hat{\alpha}_{0.5\chi_1^2+0.5\chi_2^2}$	$\hat{\alpha}_{0.5\chi_0^2+0.5\chi_1^2}$
0.01	0.016	0.049
0.05	0.055	0.174
0.10	0.103	0.311

Take home message and related works

- asymptotic distribution of LRT for general hypotheses testing
- importance of alternative hypothesis
- effect of presence of nuisance parameter

Baey, C., Cournède, P.H.,Kuhn, E.,(2019). Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models.
Computational Statistic Data Analysis

Related works

- R package VartestNlme [Baey and Kuhn (2023)]
- bootstrap test for small sample size [Guédon et al. (2024a)]
- estimating integral ratio using stochastic approximation [Guédon et al. (2024b)]
  $\rightarrow$ Tom Guédon's talk

Focus 3: Introducing genomic information in the model

Consider the following mixed effects model:

$$\begin{cases} Y_{ij} &= h(\alpha, \varphi_i, t_j) + \varepsilon_{ij} \quad 1 \leq i \leq N,\ 1 \leq j \leq J \\ \varphi_i &= \beta + b_i, \quad 1 \leq i \leq N \end{cases}$$

with $b_i \sim \mathcal{N}(0; \Gamma)$ random effects, $\varepsilon_{ij} \sim \mathcal{N}(0; \sigma^2)$ noise term

Idea : explain genotypic parameter variability with genomic information

$$\varphi_i = \mu + \beta M_i + b_i, \quad 1 \leq i \leq N$$

with M_i genomic markers of size p large versus N

$\Rightarrow$ *Variable selection in high dimension in mixed model*

Inference through reguralized maximum likelihood estimate

(on-going work, A. Caillebotte's PhD)

Consider the following mixed effects model:

$$\begin{cases} Y_{ij} = h(\alpha, \varphi_i, t_j) + \varepsilon_{ij} & 1 \leq i \leq N,\ 1 \leq j \leq J \\ \varphi_i = \mu + \beta M_i + b_i, & 1 \leq i \leq N \end{cases}$$

Consider the LASSO estimate [Tibshirani (1996)]

$$\hat{\theta}_\lambda^{LASSO} = \arg\max_{\theta \in \Theta} \left\{ \log g(\theta; y) - \lambda \|\theta\|_1 \right\}.$$

with $g(\theta; y)$ marginal likelihood and λ regularization parameter

In practice:

- Compute $\hat{\theta}_\lambda^{LASSO}$ on a grid using an adaptive stochastic weighted proximal gradient algorithm [Duchi et al. (2011)]
- Choose the regularization parameter $\hat{\lambda}$ using eBIC criterion [Chen and Chen (2009)] $\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \mathrm{eBIC}(\lambda)$

$$\mathrm{eBIC}(\lambda) = -2 \log g_\lambda(\hat{\theta}_\lambda^{MLE}; y) + |\hat{S}_\lambda| \log(NJ) + 2 \log \left(\binom{p}{|\hat{S}_\lambda|} \right)$$

Simulation study

Logistic model

$$\begin{cases} Y_{ij} = \dfrac{\varphi_{i1}}{1+\exp\left(-\dfrac{t_{ij}-\varphi_{i2}}{\alpha}\right)} + \varepsilon_{ij} & , \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2) \\ \varphi_{i1} = \mu_1 + \beta^t M_i + b_{i1} & , b_{i1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\gamma_1^2) \\ \varphi_{i2} = \mu_2 + b_{i2} & , b_{i2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\gamma_2^2) \end{cases}$$

where $M_i \in \mathbb{R}^p$ molecular markers, subject to selection, $p >> 1$

$$\theta = (\mu_1, \mu_2, \beta, \alpha, \gamma_1^2, \gamma_2^2, \sigma^2)$$

Variable selection's results

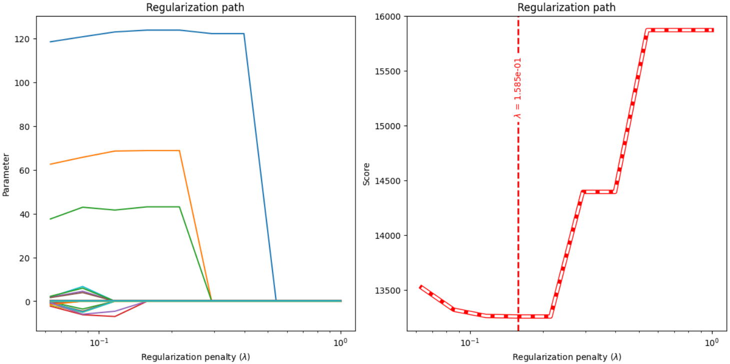
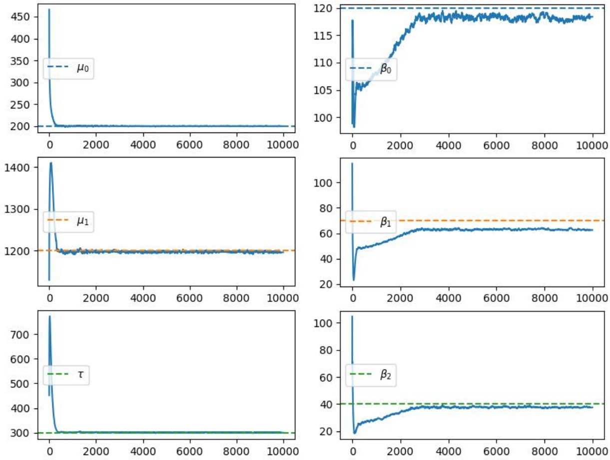


Figure: Regularization path, i.e. values of β, in solid line and the eBIC criterion in dotted line; the dotted vertical lines represent the chosen regularization values

Parameter estimates across iterations of APWSG algo



Estimates and Relative Root Mean Square Errors

		N = 100			
		P = 200		**P = 1000**	
	θ^*	$\hat{\theta}$	RRMSE	$\hat{\theta}$	RRMSE
μ_1	200.00	199.92	0.39	199.92	0.38
μ_2	1200.00	1200.23	0.36	1200.15	0.35
γ_1^2	49.00	47.71	9.29	47.56	9.13
γ_2^2	900.00	883.52	4.46	866.77	5.23
τ	300.00	300.03	0.76	300.16	0.73
σ^2	30.00	31.49	7.56	31.38	7.22
β_0	120.00	120.12	3.54	118.20	4.02
β_1	70.00	69.95	5.51	69.14	6.43
β_2	40.00	40.23	9.98	37.79	15.42

Bayesian variable selection in mixed effects models

[Naveau et al. (2024)]

Pharmacokinetic model:
$N = 200$, $p = 500$ and $J = 12$; *volume* and *dose* are known

$$\begin{cases} Y_{ij} = \frac{dose \ \varphi_{i1}}{volume \ \varphi_{i1} - \varphi_{i2}} \left(\exp(-\varphi_{i2} t_{ij}/volume) - \exp\left(-\varphi_{i1} t_{ij}\right)\right) + \varepsilon_{ij} \\ \varphi_{i1} = \mu_1 + {\beta_1}^t M_i + b_{i1} \\ \varphi_{i2} = \mu_2 + {\beta_2}^t M_i + b_{i2} \\ b_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(0, \Gamma^2) \\ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \end{cases}$$

$\Rightarrow$ compare two step approach and mixed model approach regarding robustness to partial observation settings

- complete data-set
- partial observations: only the first 3 observation times are kept for a proportion ρ of individuals, and all observation times for the remaining individuals ($\rho \in \{0.10, 0.20, 0.30, 0.40\}$)

Results for variable selection

$$Y_{ij} = \frac{dose\ \varphi_{i1}}{volume\ \varphi_{i1} - \varphi_{i2}} \left(\exp(-\varphi_{i2} t_{ij}/volume) - \exp\left(-\varphi_{i1} t_{ij}\right)\right) + \varepsilon_{ij}$$

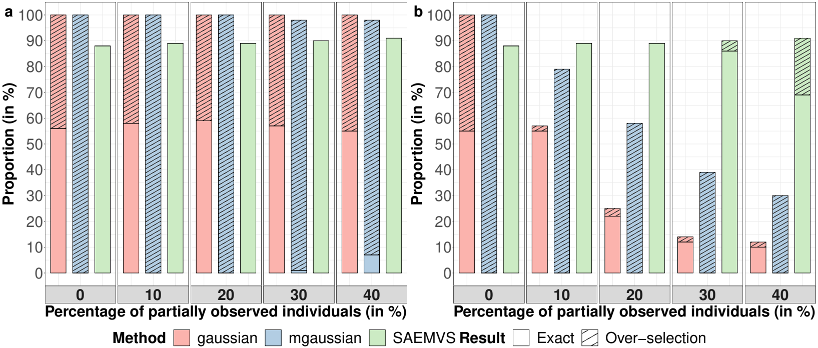


Figure: Proportion of data-sets on which the three methods select the correct model ("Exact", unpatterned bars), or a model that strictly includes the correct model ("Over-selection", striped bars) for different percentage of partially observed individuals ; left φ_1 right φ_2 .

Take home message and open questions

- modeling genotypic variability in a mechanistic model
- more interpretability
- explain variability of genotypic parameter with genomic information
- identifying relevant biomarker
- population approach regularize variable selection
- new statistical tools to reduce parameter number

$\Rightarrow$ Many open questions :

? manage correlation between covariates

? computational cost with complex mechanistic model

? post model selection inference after LASSO

? model variability within the population more finely

? ...

Bibliography

Allassonnière, S., E. Kuhn, and A. Trouvé (2007). Bayesian deformable models bulding via stochastic approximation algorithm: A convergence study. *Press in Bernoulli J*.

Baey, C., P.-H. Cournède, and E. Kuhn (2019). Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics & Data Analysis 135*, 107–122.

Baey, C., M. Delattre, E. Kuhn, J.-B. Leger, and S. Lemler (2023). Efficient preconditioned stochastic gradient descent for estimation in latent variable models. In *International Conference on Machine Learning*, pp. 1430–1453. PMLR.

Baey, C. and E. Kuhn (2023). vartestnlme: An r package for variance components testing in linear and nonlinear mixed-effects models. *Journal of Statistical Software 107*, 1–32.

Balakrishnan, S., M. J. Wainwright, and B. Yu (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in Hidden Markov Models*. Springer.

Chen, Z. and J. Chen (2009). Tournament screening cum ebic for feature selection with high-dimensional feature spaces. *Science in China Series A: Mathematics 52*(6), 1327–1341.

Davidian, M. and D. M. Giltinan (1995). *Nonlinear models for repeated measurement data*. Chapman & Hall.

Debavelaere, V. and S. Allassonnière (2021). On the curved exponential family in the stochastic approximation expectation maximization algorithm. *ESAIM: Probability & Statistics 25*.

Delattre, M. and E. Kuhn (2023). Estimating fisher information matrix in latent variable models based on the score function. *arXiv preprint arXiv:1909.06094v2*.

Delyon, B., M. Lavielle, and E. Moulines (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, 94–128.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Duchi, J., E. Hazan, and Y. Singer (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research 12*(7).

Guédon, T., C. Baey, and E. Kuhn (2024a). Bootstrap test procedure for variance components in nonlinear mixed effects models in the presence of nuisance parameters and a singular fisher information matrix. *Biometrika 111*(4), 1331–1348.

Guédon, T., C. Baey, and E. Kuhn (2024b). Estimation of ratios of normalizing constants using stochastic approximation: the saris algorithm. *arXiv preprint arXiv:2408.13022*.

Naveau, M., G. Kon Kam King, R. Rincent, L. Sansonnet, and M. Delattre (2024). Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the saem algorithm. *Statistics and Computing 34*(1), 53.

Pinheiro, J. and D. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag.

Richard-Molard, C., F. Brun, A. Laperche, M. Chelle, L. Pagès, and B. Ney (2007). Modelling n nutrition impact on plant functioning and root architecture in various genotypes of arabidopsis thaliana. In *5. International*