Variable Selection in High Dimensional Linear Mixed Model through ℓ^1 Penalization

– Jeunes Probabilistes et Satisticiens 2012 –

Florian Rohart[1,2]

April 16th-20th 2012

[1]UMR 444 Laboratoire de Génétique Cellulaire, INRA Toulouse, 31320 Castanet Tolosan cedex, France
[2]INSA, Département de Génie Mathématique, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France

Plan

Linear mixed model with q grouping factor

$$Y = X\beta + \sum_{k=1}^{q} Z_k u_k + \epsilon, \tag{1}$$

where

- Y is the set of observed data,
- X is the $n \times p$ matrix of fixed effects; $X = (X_1, \ldots, X_p)$,
- β is an unknown vector of $\mathbb{R}^p$; $\beta = (\beta_1, \ldots, \beta_p)$,
- For $k = 1, \ldots, q$, u_k is a N_k-vector of random effects, $u_k \sim \mathcal{N}(0, \sigma_k^2 I_{N_k})$,
- ϵ is a Gaussian vector with i.i.d. components $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_n)$.
- $\epsilon, u_1, \ldots, u_q$ are independent
- For $k = 1, \ldots, q$, Z_k is a $n \times N_k$ random effects regression matrix corresponding to the grouping factor k,

Linear mixed model with q grouping factor

$$Y = X\beta + \sum_{k=1}^{q} Z_k u_k + \epsilon, \tag{1}$$

where

- Y is the set of observed data,
- X is the $n \times p$ matrix of fixed effects; $X = (X_1, \ldots, X_p)$,
- β is an unknown vector of $\mathbb{R}^p$; $\beta = (\beta_1, \ldots, \beta_p)$,
- For $k = 1, \ldots, q$, u_k is a N_k-vector of random effects, $u_k \sim \mathcal{N}(0, \sigma_k^2 I_{N_k})$,
- ϵ is a Gaussian vector with i.i.d. components $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_n)$.
- $\epsilon, u_1, \ldots, u_q$ are independent
- For $k = 1, \ldots, q$, Z_k is a $n \times N_k$ random effects regression matrix corresponding to the grouping factor k,

Example: $Z_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$, $Z_2 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$

Linear mixed model with q grouping factor

$$Y = X\beta + \sum_{k=1}^{q} Z_k u_k + \epsilon, \tag{1}$$

where

- Y is the set of observed data,
- X is the $n \times p$ matrix of fixed effects; $X = (X_1, \ldots, X_p)$,
- β is an unknown vector of $\mathbb{R}^p$; $\beta = (\beta_1, \ldots, \beta_p)$,
- For $k = 1, \ldots, q$, u_k is a N_k-vector of random effects, $u_k \sim \mathcal{N}(0, \sigma_k^2 I_{N_k})$,
- ϵ is a Gaussian vector with i.i.d. components $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_n)$.
- $\epsilon, u_1, \ldots, u_q$ are independent
- For $k = 1, \ldots, q$, Z_k is a $n \times N_k$ random effects regression matrix corresponding to the grouping factor k,

$J = \{j, \beta_j \neq 0\}$. We assume that $\sum_{k=1}^{q} N_k + |J| < n$.

Aim

estimate J, β, $\sigma_1^2, \ldots, \sigma_q^2$ and σ_e^2.

Plan

Let considered β as a parameter and $\{u_k\}_{k\in\mathcal{K}}$ as missing data.
$\Phi = (\beta, \sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2)$ the vector of the parameters to estimate
$u = (u_1, \ldots, u_k)$ the missing values.
The log-likelihood of the complete data $x = (y, u)$ is

Log-likelihood

$$L(\Phi; x) = L_0(\beta, \sigma_e^2, \sigma_1^2, \ldots, \sigma_q^2; \epsilon) + \sum_{k=1}^{q} L_k(\sigma_k^2; u_k) + C, \tag{2}$$

where C is a constant of $\mathbb{R}$ and

$$-2L_0(\beta, \sigma_e^2, \sigma_u^2; \epsilon) = n ln(2\pi) + n ln(\sigma_e^2) + \left|\left| y - X\beta - \sum_{k\in\mathcal{K}} Z_k u_k \right|\right|^2 / \sigma_e^2 \tag{3a}$$

$$\text{For all } k \in \mathcal{K}, -2L_k(\sigma_k^2; u_k) = N_k ln(2\pi) + N_k ln(\sigma_k^2) + ||u_k||^2 / \sigma_k^2 \tag{3b}$$

Let considered β as a parameter and $\{u_k\}_{k\in\mathcal{K}}$ as missing data.
$\Phi = (\beta, \sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2)$ the vector of the parameters to estimate
$u = (u_1, \ldots, u_k)$ the missing values.
The log-likelihood of the complete data $x = (y, u)$ is

Log-likelihood

$$L(\Phi; x) = L_0(\beta, \sigma_e^2, \sigma_1^2, \ldots, \sigma_q^2; \epsilon) + \sum_{k=1}^{q} L_k(\sigma_k^2; u_k) + C, \tag{2}$$

where C is a constant of $\mathbb{R}$ and

$$-2L_0(\beta, \sigma_e^2, \sigma_u^2; \epsilon) = n ln(2\pi) + n ln(\sigma_e^2) + \left|\left| y - X\beta - \sum_{k\in\mathcal{K}} Z_k u_k \right|\right|^2 / \sigma_e^2 \tag{3a}$$

$$\text{For all } k \in \mathcal{K}, -2L_k(\sigma_k^2; u_k) = N_k ln(2\pi) + N_k ln(\sigma_k^2) + ||u_k||^2 / \sigma_k^2 \tag{3b}$$

Objective function

$$g(\Phi; x) = -2L(\Phi; x) + \lambda|\beta|_1 \tag{4}$$

Problems

- non-linear, non-differentiable and non convex problem

Problems

- non-linear, non-differentiable and non convex problem
- Let $k \in \mathcal{K}$, the function

$$-2L_k(\sigma_k^2; u_k) = N_k ln(2\pi) + N_k ln(\sigma_k^2) + ||u_k||^2 / \sigma_k^2$$

is **NOT** lower-bounded....

$$\lim_{(\sigma_k^2, u_k) \to (0, 0_{N_k})} -2L_k = -\infty$$

Problems

- non-linear, non-differentiable and non convex problem
- Let $k \in \mathcal{K}$, the function

$$-2L_k(\sigma_k^2; u_k) = N_k ln(2\pi) + N_k ln(\sigma_k^2) + ||u_k||^2 / \sigma_k^2$$

is **NOT** lower-bounded....

$$\lim_{(\sigma_k^2, u_k) \to (0, 0_{N_k})} -2L_k = -\infty$$

Well-known problem for gaussian mixture of degeneracy of the likelihood (Biernacki and Chrétien, 2003) but not much concerning mixed model because algorithms usually focus on the log-likelihood of the model (Schelldorfer et al., 2011):

$$Y = X\beta + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, V), V = ZGZ' + R \tag{5}$$

Taking advantage of the problems: Selection of random effects

Selection of random effects

When the minimization of g coincides with the annulation of the random effect k, this random effect is deleted from the model

Taking advantage of the problems: Selection of random effects

Selection of random effects

When the minimization of g coincides with the annulation of the random effect k, this random effect is deleted from the model

This can happen for two reasons:

- the true underlying model was different from the fitted one
- the initialization of the convergence process was to close to an attraction domain of $(u_k, \sigma_k^2) = (0_{N_k}, 0)$ (Biernacki and Chrétien, 2003)

Taking advantage of the problems: Selection of random effects

Selection of random effects

When the minimization of g coincides with the annulation of the random effect k, this random effect is deleted from the model

This can happen for two reasons:

- the true underlying model was different from the fitted one
- the initialization of the convergence process was to close to an attraction domain of $(u_k, \sigma_k^2) = (0_{N_k}, 0)$ (Biernacki and Chrétien, 2003)

Remark: When a random effect is deleted from the model with q grouping factor, the objective function is modified accordingly.
It can go on until no grouping factor remains, then a linear model is considered.

Plan

A multicycle ECM

Algorithm 1 Initialization:

Initialize the set of parameters $\Phi^{[0]} = (\sigma_{\mathcal{K}}^{2[0]}, \sigma_e^{2[0]}, \beta^{[0]})$. Set $\mathcal{K} = \{1, \ldots, q\}$.

Define $\Gamma^{[0]}$ as the block diagonal matrix of $\gamma_1^{[0]} I_{N_1}, \ldots, \gamma_q^{[0]} I_{N_q}$, where $\gamma_k^{[0]} = \sigma_e^{2[0]} / \sigma_k^{2[0]}$.

Until convergence:

1. *E-step*

$u^{[t+1/2]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t]})$

2. *M-step*

$$\beta^{[t+1]} = \underset{\beta}{Argmin} \left(\left|\left| \left(y - Z\hat{u}^{[t+1/2]}\right) - X\beta \right|\right|^2 + \lambda * \sigma_e^{2[t]} |\beta|_1 \right)$$

3. *E-step*

$u^{[t+1]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t+1]})$

4. *M-step*

(a) For k in $\mathcal{K}$

Set $\sigma_k^{2[t+1]} = \left|\left| [u^{[t+1]}]_k \right|\right|^2 / N_k + tr\left(T_{k,k}\right) \sigma_e^{2[t]} / N_k$

if $\left(\left|\left| [u^{[t+1]}]_k \right|\right|^2 / N_k < 10^{-6} \right)$ then $\mathcal{K} = \mathcal{K}_k$

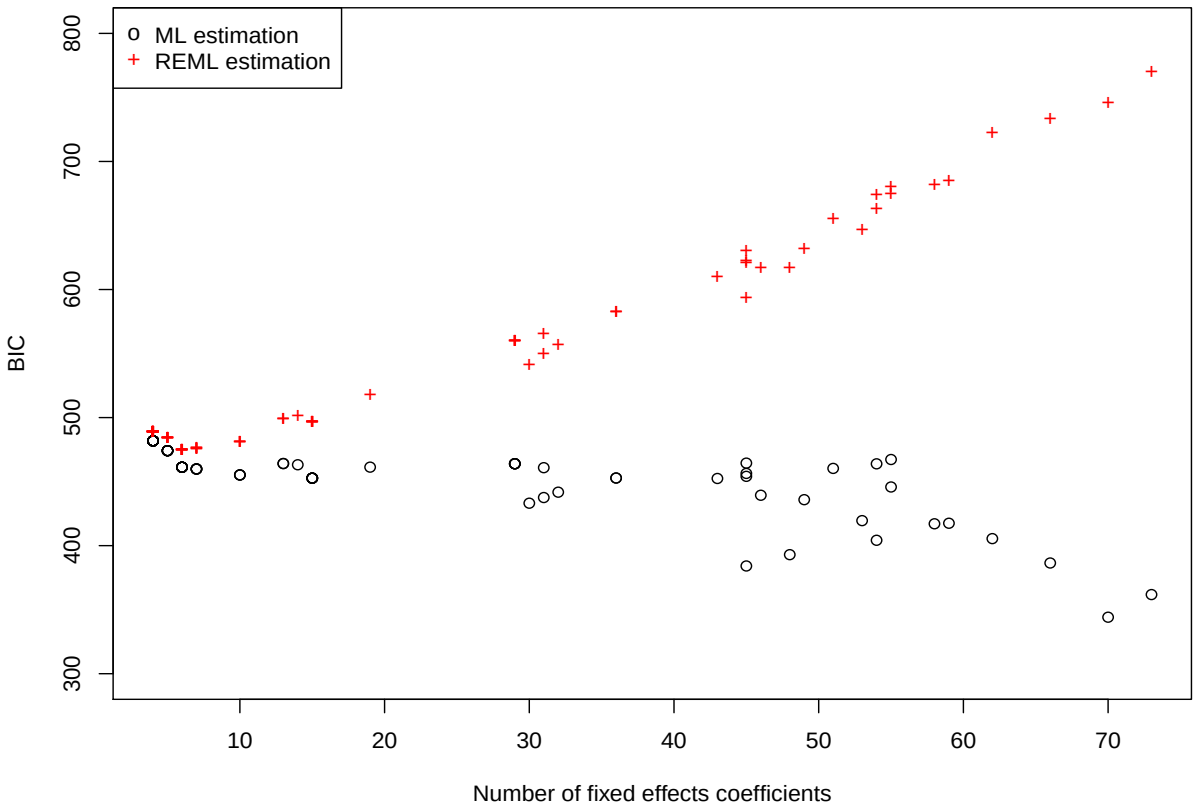
(b) Set $\sigma_e^{2(t+1)} = \frac{1}{n} \left[\left|\left| y - X\beta^{[t+1]} - Z\hat{u}^{[t+1]} \right|\right|^2 + \sum_{k=1}^{q} \left(N_k - \gamma_k^{[t]} tr\left(T_{k,k}\right) \sigma_e^{2[t]} \right) \right]$

Set $\Gamma^{[t+1]}$ as the block diagonal matrix of $\gamma_1^{[t+1]} I_{N_1}, \ldots, \gamma_q^{[t+1]} I_{N_q}$,

where $\gamma_k^{[t+1]} = \sigma_e^{2[t+1]} / \sigma_k^{2[t+1]}$.

BIC with ML estimation can end up with a clear overestimation of the set of indices of the relevant fixed effects

The Bayesian Information Criterion (BIC) is used from a REML estimate on the model that has been selected with λ.



Plan

We considered that the variances are known.
$\gamma_k = \sigma_e^2/\sigma_k^2$, Γ is the block diagonal matrix of $\gamma_1 I_{N_1}, \ldots, \gamma_q I_{N_q}$, $Z = (Z_1, \ldots, Z_q)$

Objective function

$$g_{G,R}(\beta, u_1, \ldots, u_q) = ||Y - X\beta - \sum_{k=1}^{q} Z_k u_k||_2^2 + \sum_{k=1}^{q} \frac{\sigma_e^2}{\sigma_k^2} ||u_k||_2^2 + \lambda|\beta|_1, \qquad (6)$$

Convex problem $->$ unique minimum.

We considered that the variances are known.
$\gamma_k = \sigma_e^2/\sigma_k^2$, Γ is the block diagonal matrix of $\gamma_1 I_{N_1}, \ldots, \gamma_q I_{N_q}$, $Z = (Z_1, \ldots, Z_q)$

Objective function

$$g_{G,R}(\beta, u_1, \ldots, u_q) = ||Y - X\beta - \sum_{k=1}^{q} Z_k u_k||_2^2 + \sum_{k=1}^{q} \frac{\sigma_e^2}{\sigma_k^2} ||u_k||_2^2 + \lambda|\beta|_1, \qquad (6)$$

Convex problem $->$ unique minimum.

Algorithm Initialization: Initialize $\beta^{[0]}$.
Until convergence:

1. *E-step*
$u^{[t+1]} = (Z'Z + \Gamma)^{-1} Z' \left(y - X\beta^{[t]}\right)$
2. *M-step*
$\beta^{[t+1]} = \underset{\beta}{Argmin} \left(\left|\left| \left(y - Z\hat{u}^{[t+1]}\right) - X\beta \right|\right|^2 + \lambda |\beta|_1 \right)$

Back to Linear Model

Let $\tilde{Y} = (Y, 0_N)'$ where $N = \sum_{k=1}^{q} N_k$ and let $b = \left(u_1', \ldots, u_q', \beta'\right)'$.
Set $A = \begin{pmatrix} Z & X \\ \sqrt{\Gamma} & 0_{N \times p} \end{pmatrix}$. With these notations, we have:

Objective function

$$g_{G,R}(\beta, u_1, \ldots, u_q) = ||\tilde{Y} - Ab||_2^2 + \lambda \sum_{k=N+1}^{N+p} |b_k| . \tag{7}$$

Back to Linear Model

Let $\tilde{Y} = (Y, 0_N)'$ where $N = \sum_{k=1}^{q} N_k$ and let $b = \left(u_1', \ldots, u_q', \beta'\right)'$.
Set $A = \begin{pmatrix} Z & X \\ \sqrt{\Gamma} & 0_{N \times p} \end{pmatrix}$. With these notations, we have:

Objective function

$$g_{G,R}(\beta, u_1, \ldots, u_q) = ||\tilde{Y} - Ab||_2^2 + \lambda \sum_{k=N+1}^{N+p} |b_k| . \tag{7}$$

Thus, the minimization of (6) in $\beta, u_1, \ldots, u_q$ can be viewed as the minimization of (7) in b, which is a Lasso in the linear model:

$$\tilde{Y} = Ab + \epsilon, \tag{8}$$

where ϵ is a Gaussian vector with i.i.d. components of known variance σ_e^2.

Any variable selection method can be used to estimate b with the theoretical results that go along!!

Plan

Algorithm 2 Initialization:

Initialize the set of parameters $\Phi^{[0]} = (\sigma_{\mathcal{K}}^{2[0]}, \sigma_e^{2[0]}, \beta^{[0]})$. Set $\mathcal{K} = \{1, \ldots, q\}$.

Define $\Gamma^{[0]}$ as the block diagonal matrix of $\gamma_1^{[0]} I_{N_1}, \ldots, \gamma_q^{[0]} I_{N_q}$, where $\gamma_k^{[0]} = \sigma_e^{2[0]} / \sigma_k^{2[0]}$.

Until convergence:

1. $u^{[t+1/2]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t]})$
2. $\beta^{[t+1]} = \underset{\beta}{Argmin} \left(\left|\left| (y - Z\hat{u}^{[t+1/2]}) - X\beta \right|\right|^2 + \lambda * \sigma_e^{2[t]} |\beta|_1 \right)$
3. $u^{[t+1]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t+1]})$
4. (a) For k in $\mathcal{K}$
   - Set $\sigma_k^{2[t+1]} = \left|\left| [u^{[t+1]}]_k \right|\right|^2 / N_k + tr\left(T_{k,k}\right) \sigma_e^{2[t]} / N_k$
   - if $\left(\left|\left| [u^{[t+1]}]_k \right|\right|^2 / N_k < 10^{-6} \right)$ then $\mathcal{K} = \mathcal{K}_k$

   (b) Set $\sigma_e^{2(t+1)} = \frac{1}{n} \left[\left|\left| y - X\beta^{[t+1]} - Z\hat{u}^{[t+1]} \right|\right|^2 + \sum_{k=1}^{q} \left(N_k - \gamma_k^{[t]} tr\left(T_{k,k}\right) \sigma_e^{2[t]} \right) \right]$

   Set $\Gamma^{[t+1]}$ as the block diagonal matrix of $\gamma_1^{[t+1]} I_{N_1}, \ldots, \gamma_q^{[t+1]} I_{N_q}$, where $\gamma_k^{[t+1]} = \sigma_e^{2[t+1]} / \sigma_k^{2[t+1]}$

Algorithm 2 Initialization:

Initialize the set of parameters $\Phi^{[0]} = (\sigma_{\mathcal{K}}^{2[0]}, \sigma_e^{2[0]}, \beta^{[0]})$. Set $\mathcal{K} = \{1, \ldots, q\}$.

Define $\Gamma^{[0]}$ as the block diagonal matrix of $\gamma_1^{[0]} I_{N_1}, \ldots, \gamma_q^{[0]} I_{N_q}$, where $\gamma_k^{[0]} = \sigma_e^{2[0]} / \sigma_k^{2[0]}$.

Until convergence:

1. $u^{[t+1/2]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t]})$
2. **Variable selection and estimation of β in the linear model $Y - Zu^{[t+1/2]} = X\beta + \epsilon^{[t]}$, where $\epsilon^{[t]}$ is a i.i.d Gaussian noise.**
3. $u^{[t+1]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t+1]})$
4. (a) For k in $\mathcal{K}$
   - Set $\sigma_k^{2[t+1]} = \left|\left|[u^{[t+1]}]_k\right|\right|^2 / N_k + tr\left(T_{k,k}\right) \sigma_e^{2[t]} / N_k$
   - if $\left(\left|\left|[u^{[t+1]}]_k\right|\right|^2 / N_k < 10^{-6}\right)$ then $\mathcal{K} = \mathcal{K}_k$

   (b) Set $\sigma_e^{2(t+1)} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Z\hat{u}^{[t+1]}\right|\right|^2 + \sum_{k=1}^{q}\left(N_k - \gamma_k^{[t]} tr\left(T_{k,k}\right)\sigma_e^{2[t]}\right)\right]$

   Set $\Gamma^{[t+1]}$ as the block diagonal matrix of $\gamma_1^{[t+1]} I_{N_1}, \ldots, \gamma_q^{[t+1]} I_{N_q}$, where $\gamma_k^{[t+1]} = \sigma_e^{2[t+1]} / \sigma_k^{2[t+1]}$

Plan

Alg1 and lmmlasso gives similar results, Alg2 combined with multiple testing outperformed them

	Ideal	lasso	lasso+	lmmlasso	procbol	procbol+
Truth	1	0.00	0.31	0.33	0.15	0.82
$\|\hat{J}\|$	5	23.45(17.79)	6.29(1.25)	6.35(1.34)	3.85(1.00)	5.21(0.56)
TP	5	4.06(0.98)	4.99(0.10)	4.99(0.10)	3.61(0.95)	4.99(0.10)
$\hat{\sigma}_e^2$		1.54(0.97)	1.33(0.26)	1.32(0.25)	3.23(0.74)	0.92(0.17)
$\hat{\sigma}_1^2$	1	-	0.92(0.41)	0.92(0.41)	-	0.97(0.39)
$\hat{\sigma}_2^2$	1	-	1.03(0.41)	1.04(0.40)	-	1.00(0.34)
$\hat{\beta}_1$	0.67	0.60(0.25)	0.62(0.25)	0.62(0.25)	0.60(0.25)	0.60(0.25)
$\hat{\beta}_2$	0.67	0.28(0.32)	0.54(0.27)	0.55(0.27)	0.55(0.31)	0.64(0.28)
$\hat{\beta}_3$	0.67	0.40(0.24)	0.27(0.11)	0.28(0.11)	0.38(0.38)	0.67(0.10)
$\hat{\beta}_4$	0.67	0.43(0.24)	0.31(0.12)	0.31(0.12)	0.44(0.39)	0.67(0.10)
$\hat{\beta}_5$	0.67	0.46(0.24)	0.34(0.12)	0.34(0.12)	0.43(0.40)	0.67(0.13)
mse	0	1.84(0.67)	0.52(0.19)	0.52(0.19)	0.83(0.43)	0.21(0.15)

Table: $n = 120$, $p = 600$, $|J| = 5$, $\beta_J = 2/3$, $q = 2$, $SNR = 0.63(0.11)$, the two grouping factor are equals

lmmlasso, see Schelldorfer et al. (2011); lasso+: Algorithm 1;
procbol: multiple hypotheses testing (Rohart, 2011); pbol+: Algorithm 2 combined with procbol (mht-package in R)

Alg1 and lmmlasso gives similar results, Alg2 combined with multiple testing outperformed them

	Ideal	lasso	lasso+	lmmlasso	procbol	procbol+
Truth	1	0.00	0.31	0.33	0.15	0.82
$\lvert\hat{J}\rvert$	5	23.45(17.79)	6.29(1.25)	6.35(1.34)	3.85(1.00)	5.21(0.56)
TP	5	4.06(0.98)	4.99(0.10)	4.99(0.10)	3.61(0.95)	4.99(0.10)
$\hat{\sigma}^2_e$		1.54(0.97)	1.33(0.26)	1.32(0.25)	3.23(0.74)	0.92(0.17)
$\hat{\sigma}^2_1$	1	-	0.92(0.41)	0.92(0.41)	-	0.97(0.39)
$\hat{\sigma}^2_2$	1	-	1.03(0.41)	1.04(0.40)	-	1.00(0.34)
$\hat{\beta}_1$	0.67	0.60(0.25)	0.62(0.25)	0.62(0.25)	0.60(0.25)	0.60(0.25)
$\hat{\beta}_2$	0.67	0.28(0.32)	0.54(0.27)	0.55(0.27)	0.55(0.31)	0.64(0.28)
$\hat{\beta}_3$	0.67	0.40(0.24)	0.27(0.11)	0.28(0.11)	0.38(0.38)	0.67(0.10)
$\hat{\beta}_4$	0.67	0.43(0.24)	0.31(0.12)	0.31(0.12)	0.44(0.39)	0.67(0.10)
$\hat{\beta}_5$	0.67	0.46(0.24)	0.34(0.12)	0.34(0.12)	0.43(0.40)	0.67(0.13)
mse	0	1.84(0.67)	0.52(0.19)	0.52(0.19)	0.83(0.43)	0.21(0.15)

Table: $n = 120$, $p = 600$, $|J| = 5$, $\beta_J = 2/3$, $q = 2$, $SNR = 0.63(0.11)$, the two grouping factor are equals

lmmlasso, see Schelldorfer et al. (2011); lasso+: Algorithm 1;
procbol: multiple hypotheses testing (Rohart, 2011); pbol+: Algorithm 2 combined with procbol (mht-package in R)

Alg1 and lmmlasso gives similar results, Alg2 combined with multiple testing outperformed them

	Ideal	lasso	lasso+	lmmlasso	procbol	procbol+
Truth	1	0.00	0.31	0.33	0.15	0.82
$\lvert\hat{J}\rvert$	5	23.45(17.79)	6.29(1.25)	6.35(1.34)	3.85(1.00)	5.21(0.56)
TP	5	4.06(0.98)	4.99(0.10)	4.99(0.10)	3.61(0.95)	4.99(0.10)
$\hat{\sigma}_e^2$		1.54(0.97)	1.33(0.26)	1.32(0.25)	3.23(0.74)	0.92(0.17)
$\hat{\sigma}_1^2$	1	-	0.92(0.41)	0.92(0.41)	-	0.97(0.39)
$\hat{\sigma}_2^2$	1	-	1.03(0.41)	1.04(0.40)	-	1.00(0.34)
$\hat{\beta}_1$	0.67	0.60(0.25)	0.62(0.25)	0.62(0.25)	0.60(0.25)	0.60(0.25)
$\hat{\beta}_2$	0.67	0.28(0.32)	0.54(0.27)	0.55(0.27)	0.55(0.31)	0.64(0.28)
$\hat{\beta}_3$	0.67	0.40(0.24)	0.27(0.11)	0.28(0.11)	0.38(0.38)	0.67(0.10)
$\hat{\beta}_4$	0.67	0.43(0.24)	0.31(0.12)	0.31(0.12)	0.44(0.39)	0.67(0.10)
$\hat{\beta}_5$	0.67	0.46(0.24)	0.34(0.12)	0.34(0.12)	0.43(0.40)	0.67(0.13)
mse	0	1.84(0.67)	0.52(0.19)	0.52(0.19)	0.83(0.43)	0.21(0.15)

Table: $n = 120$, $p = 600$, $|J| = 5$, $\beta_J = 2/3$, $q = 2$, $SNR = 0.63(0.11)$, the two grouping factor are equals

lmmlasso, see Schelldorfer et al. (2011); lasso+: Algorithm 1;
procbol: multiple hypotheses testing (Rohart, 2011); pbol+: Algorithm 2 combined with procbol (mht-package in R)

Alg1 and lmmlasso gives similar results, Alg2 combined with multiple testing outperformed them

	Ideal	lasso	lasso+	lmmlasso	procbol	procbol+
Truth	1	0.00	0.31	0.33	0.15	0.82
$\lvert\hat{J}\rvert$	5	23.45(17.79)	6.29(1.25)	6.35(1.34)	3.85(1.00)	5.21(0.56)
TP	5	4.06(0.98)	4.99(0.10)	4.99(0.10)	3.61(0.95)	4.99(0.10)
$\hat{\sigma}_e^2$		1.54(0.97)	1.33(0.26)	1.32(0.25)	3.23(0.74)	0.92(0.17)
$\hat{\sigma}_1^2$	1	-	0.92(0.41)	0.92(0.41)	-	0.97(0.39)
$\hat{\sigma}_2^2$	1	-	1.03(0.41)	1.04(0.40)	-	1.00(0.34)
$\hat{\beta}_1$	0.67	0.60(0.25)	0.62(0.25)	0.62(0.25)	0.60(0.25)	0.60(0.25)
$\hat{\beta}_2$	0.67	0.28(0.32)	0.54(0.27)	0.55(0.27)	0.55(0.31)	0.64(0.28)
$\hat{\beta}_3$	0.67	0.40(0.24)	0.27(0.11)	0.28(0.11)	0.38(0.38)	0.67(0.10)
$\hat{\beta}_4$	0.67	0.43(0.24)	0.31(0.12)	0.31(0.12)	0.44(0.39)	0.67(0.10)
$\hat{\beta}_5$	0.67	0.46(0.24)	0.34(0.12)	0.34(0.12)	0.43(0.40)	0.67(0.13)
mse	0	1.84(0.67)	0.52(0.19)	0.52(0.19)	0.83(0.43)	0.21(0.15)

Table: $n = 120$, $p = 600$, $|J| = 5$, $\beta_J = 2/3$, $q = 2$, $SNR = 0.63(0.11)$, the two grouping factor are equals

lmmlasso, see Schelldorfer et al. (2011); lasso+: Algorithm 1;
procbol: multiple hypotheses testing (Rohart, 2011); pbol+: Algorithm 2 combined with procbol (mht-package in R)

REML estimation on the selected model

	Ideal	lmmlasso	lasso+
$\hat{\sigma}_e^2$	1	0.92(0.09)	0.93(0.09)
$\hat{\sigma}_1^2$	1	1.02(0.41)	1.02(0.42)
$\hat{\sigma}_2^2$	1	1.08(0.37)	1.07(0.38)
$\hat{\beta}_1$	0.67	0.61(0.25)	0.61(0.25)
$\hat{\beta}_2$	0.67	0.62(0.27)	0.62(0.28)
$\hat{\beta}_3$	0.67	0.63(0.11)	0.63(0.11)
$\hat{\beta}_4$	0.67	0.64(0.12)	0.64(0.12)
$\hat{\beta}_5$	0.67	0.64(0.14)	0.64(0.13)
mse	0	0.31(0.17)	0.30(0.17)

Biernacki, C. and Chrétien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with em. *Statistics & Probability Letters*, 61:373–382.

Foulley, J. (1997). Ecm approaches to heteroskedastic mixed models with constant variance ratios. *Genetics Selection Evolution*, 29:197–318.

Henderson, C. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, pages 10–41.

McLachlan, J. and T., K. (2008). *The EM Algorithm and Extensions, second edition*. Wiley-Interscience.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80:267–278.

Rohart, F. (2011). Multiple hypotheses testing for variable selection. *arXiv:1106.3415v1*.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38:197–214.

Thank you!?

Plan

It is a multicycle ECM algorithm (Foulley, 1997; McLachlan and T., 2008; Meng and Rubin, 1993)

- $\Phi = (\beta, \sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2)$ is the vector of the parameters to estimate
- $u = (u_1, \ldots, u_k)$ is the vector of the missing values
- Z denotes the concatenation of $(Z_1, \ldots, Z_q)$
- Γ denotes the block diagonal matrix of $\gamma_1 I_{N_1}, \ldots, \gamma_q I_{N_q}$, where $\gamma_k = \sigma_e^2/\sigma_k^2$.
- $\mathcal{K} = \{1, \ldots, q\}$, $\mathcal{K}_j = \mathcal{K} \backslash \{j\}$

Let denote

$$Q(\Phi; \Phi^{[t]}) = E_{u|y,\Phi=\Phi^{[t]}}[g(\Phi; x)] \tag{9}$$

We can decomposed Q as following:

$$Q(\Phi; \Phi^{[t]}) = Q_0(\beta, \sigma_{\mathcal{K}}^2, \sigma_e^2; \Phi^{[t]}) + \sum_{k=1}^{q} Q_k(\sigma_k^2; \Phi^{[t]}) \tag{10}$$

with

$$Q_0(\Phi; \Phi^{[t]}) = nln(2\pi) + nln(\sigma_e^2) + E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon)/\sigma_e^2 + \lambda|\beta|_1 \tag{11a}$$

$$\forall k \in \mathcal{K}, Q_k(\sigma_k^2; \Phi^{[t]}) = Nln(2\pi) + Nln(\sigma_k^2) + E_{u|y,\Phi=\Phi^{[t]}}(u_k{}'u_k)/\sigma_k^2 \tag{11b}$$

By definition, we have

$$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon) = \left|\left|E_{u|y,\Phi=\Phi^{[t]}}(\epsilon)\right|\right|^2 + tr\left(Var_{u|y,\Phi=\Phi^{(t)}}(\epsilon)\right) \tag{12}$$

Thus,

$$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon) = \left|\left|y - X\beta^{[t]} - ZE\left(u|y,\Phi=\Phi^{[t]}\right)\right|\right|^2 + tr\left(Zvar\left(u|y,\Phi^{[t]}\right)Z'\right). \tag{13}$$

According to the denomination of Henderson (1973), $E\left(u|y,\Phi=\Phi^{(t)}\right)$ is the BLUP (Best Linear Unbiased Prediction) of u. Let denote $\hat{u}^{[t+1/2]} = E\left(u|y,\Phi=\Phi^{(t)}\right)$, thus

E-Step

$$\hat{u}^{[t+1/2]} = (Z'Z + \Gamma^{[t]})^{-1}Z'\left(y - X\beta^{[t]}\right) \tag{14}$$

The next step performs a minimization of $Q_0(\beta, \sigma^2_{\mathcal{K}}, \sigma^2_e; \Phi^{[t]})$ with respect to β:

M-Step for β

$$\beta^{[t+1]} = \underset{\beta}{Argmin} \left(\frac{1}{\sigma_e^{2[t]}} \left|\left| \left(y - Z\hat{u}^{[t+1/2]}\right) - X\beta \right|\right|^2 + \lambda \left|\beta\right|_1 \right). \tag{15}$$

Remark that (15) is a Lasso on β with the vector of "observed" data $\left(y - Z\hat{u}^{[t+1/2]}\right)$ and the penalty $\lambda * \sigma_e^{2[t]}$.

A second E-step is performed with the actualization of the vector of missing values u:

E-Step

$$\hat{u}^{[t+1]} = (Z'Z + \Gamma^{[t]})^{-1} Z' \left(y - X\beta^{[t+1]} \right). \tag{16}$$

We define $\forall k \in \mathcal{K}$, $\hat{u}_k^{[t+1]} := [\hat{u}^{[t+1]}]_k$ to be the element of size N_k that corresponds to the grouping factor k in $\hat{u}^{[t+1]}$.

Let $k \in \mathcal{K}$. The minimization of Q_1 with respect to σ_k^2 gives:

$$\sigma_k^{2[t+1]} = E\left(u_k' u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right) / N_k$$

Besides,

$$E\left(u_k' u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right) = || \underbrace{E\left(u_k | -\right)}_{=\hat{u}_k^{[t+1]}} ||^2 + tr\left(var\left(u_k | -\right)\right). \tag{17}$$

Moreover,

$$var\left(u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right) = T_{k,k} \sigma_e^{2[t]}, \tag{18}$$

where $T_{k,k}$ is defined as following:

$$\left(Z'Z + \Gamma^{[t]}\right)^{-1} = \begin{pmatrix} Z_1'Z_1 + \gamma_1^{[t]} I_{N_1} & Z_1'Z_2 & \cdots & Z_1'Z_q \\ Z_2'Z_1 & Z_2'Z_2 + \gamma_2^{[t]} I_{N_2} & \cdots & Z_2'Z_q \\ \vdots & \vdots & \ddots & \vdots \\ Z_q'Z_1 & Z_q'Z_2 & \cdots & Z_q'Z_q + \gamma_q^{[t]} I_{N_q} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,q} \\ T_{1,2}' & T_{2,2} & \cdots & T_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ T_{1,q}' & T_{2,q}' & \cdots & T_{q,q} \end{pmatrix}.$$

$\sigma_k^{2[t+1]}$

Thus, for all $k \in \mathcal{K}$:

$$\sigma_k^{2[t+1]} = \frac{1}{N_k} \left[\left|\left| \hat{u}_k^{[t+1]} \right|\right|^2 + tr\left(T_{k,k}\right) \sigma_e^{2[t]} \right] \tag{19}$$

The minimization of Q_0 with respect to σ_e^2 gives:

$$\sigma_e^{2[t+1]} = E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon)/n$$

From (13):

$$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon) = \left|\left|y - X\beta^{[t]} - ZE\left(u|y,\Phi=\Phi^{[t]}\right)\right|\right|^2 + tr\left(Zvar\left(u|y,\Phi^{[t]}\right)Z'\right),$$

we have

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Z\hat{u}^{[t+1]}\right|\right|^2 + tr\left(Z(Z'Z+\Gamma^{[t]})^{(-1)}Z'\right)\sigma_e^{2[t]}\right]. \quad (20)$$

Since

$$tr\left(Z\left(Z'Z+\Gamma^{(t)}\right)^{(-1)}Z'\right) = N - \sum_{k=1}^{q} tr\left(T_{k,k}\right)\gamma_k^{[t]} \quad (21)$$

we have

$\sigma_e^{2[t+1]}$

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Z\hat{u}^{[t+1]}\right|\right|^2 + \left(N - \sum_{k=1}^{q}\gamma_k^{[t]}tr\left(T_{k,k}\right)\right)\sigma_e^{2[t]}\right] \quad (22)$$