

Extension of the EM-algorithm using PLS to fit linear mixed effects models for high dimensional repeated data

Caroline Bazzoli, Sophie Lambert-Lacroix, Marie-José Martinez

► **To cite this version:**

Caroline Bazzoli, Sophie Lambert-Lacroix, Marie-José Martinez. Extension of the EM-algorithm using PLS to fit linear mixed effects models for high dimensional repeated data . IBC 2018-29th International Biometric Conference, Jul 2018, Barcelona, Spain. hal-01834609

HAL Id: hal-01834609

<https://hal.archives-ouvertes.fr/hal-01834609>

Submitted on 10 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

EXTENSION OF THE EM-ALGORITHM USING PLS TO FIT LINEAR MIXED EFFECTS MODELS FOR HIGH DIMENSIONAL REPEATED DATA



Caroline Bazzoli⁽¹⁾, Sophie Lambert-Lacroix⁽²⁾, Marie-José Martinez⁽¹⁾

⁽¹⁾ Laboratoire Jean Kuntzmann, Université Grenoble Alpes, Grenoble, FRANCE

⁽²⁾ TIMC-IMAG, Université Grenoble Alpes, La Tronche, FRANCE



I. Introduction

To deal with repeated data

- Linear mixed effects models are highly recommended.
- A classical parameter estimation method: Expectation- Maximization (EM) algorithm.

To deal with high-dimensional data

- Reduction dimension methods can be used to summarize the numerous predictors in form of a small number of new components.
- Classical approach : Principal Component Regression (PCR)
 - Does not consider the link between the outcome and the independent variables.
- Alternative method: Partial Least Squares (PLS)
 - Takes the link between the outcome and the independent variables into account.

To solve the high dimensional issue in the repeated data context

→ Introduction of a PLS step into the EM-algorithm for linear mixed models to reduce the high-dimensional data.

- **Idea:** At each iteration, the outcome data is substituted in the input of PLS by a pseudo-variable response whose expected value has a linear relationship with the covariates.

III. Extension of the EM-algorithm for high dimensional repeated data

Extension of the EM-algorithm using PLS (EM-PLS)

At iteration $[t + 1]$:

- E-step: Compute the expectation of the complete data log-likelihood given the observed data and a current value of the parameters $\theta^{[t]}$
- M-step:
 - Define the pseudo-response variable $z^{[t+1]} = X\beta^{[t]} + \sigma^{2[t]} \Gamma^{[t]-1} (y - X\beta^{[t]})$
 - Perform PLS regression of the pseudo-response variable $z^{[t+1]}$ onto X :

$$\beta^{[t+1]} \leftarrow PLS(z^{[t+1]}, X, \kappa)$$

where κ is the PLS component number obtained by cross-validation.

- Calculate new parameter values $\tau^{2[t+1]}$ and $\sigma^{2[t+1]}$ from Equations (2) and (3).

Extension of the EM-algorithm using PCR (EM-PCR)

Similarly, a PCR step is introduced into the EM-algorithm to reduce the high-dimensional data to low-dimensional features.

II. The linear mixed effects model

Definition of the linear mixed effects model

- $Y = (Y_1', \dots, Y_I')'$ with $Y_i = (Y_{i1}, \dots, Y_{in_i})'$ the vector of all measurements for the i th individual, $i = 1, \dots, I$.
- The linear mixed model for the response Y is defined as

$$Y = X\beta + U\xi + \varepsilon$$

with X the $n \times p$ design matrix associated to the p -vector fixed effects β , U the $n \times I$ design vector associated to the random effects $\xi = (\xi_1, \dots, \xi_I)'$, $\xi \sim \mathcal{N}(0_I, \tau^2 I_I)$ and $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.

- The marginal distribution of the response Y is given by

$$Y \sim \mathcal{N}(X\beta, \Gamma) \quad \text{with} \quad \Gamma = \tau^2 U U' + \sigma^2 I_n$$

The ML estimation approach using the EM-algorithm

- The log-likelihood associated with the complete data (y, ξ) is given by

$$L(\theta|y, \xi) = -\frac{1}{2} \left\{ (n+I) \ln 2\pi + I \ln \tau^2 + n \ln \sigma^2 + \frac{(y - X\beta - U\xi)'(y - X\beta - U\xi)}{\tau^2} + \frac{\xi^2}{\sigma^2} \right\}$$

- At iteration $[t + 1]$, the E-step consists of computing the expectation of the complete data log-likelihood given the observed data and a current value of the parameters $\theta^{[t]} = (\beta^{[t]}, \tau^{2[t]}, \sigma^{2[t]})$:

$$Q(\theta|\theta^{[t]}) = E[L(\theta|y, \xi)|y, \theta^{[t]}]$$

- The M-step consists of maximizing $Q(\theta|\theta^{[t]})$. It leads to the following explicit expressions:

$$\beta^{[t+1]} = (X'X)^{-1} X' \left\{ X\beta^{[t]} + \sigma^{2[t]} \Gamma^{[t]-1} (y - X\beta^{[t]}) \right\} \quad (1)$$

$$\tau^{2[t+1]} = \frac{1}{I} \left\{ \tau^{4[t]} (y - X\beta^{[t]})' \Gamma^{[t]-1} U U' \Gamma^{[t]-1} (y - X\beta^{[t]}) + \tau^{2[t]} - \tau^{4[t]} \text{tr}(\Gamma^{[t]-1} U U') \right\} \quad (2)$$

$$\sigma^{2[t+1]} = \frac{1}{n} \left\{ \sigma^{4[t]} (y - X\beta^{[t]})' \Gamma^{[t]-1} \Gamma^{[t]-1} (y - X\beta^{[t]}) + n \sigma^{2[t]} - \tau^{4[t]} \text{tr}(\Gamma^{[t]-1}) \right\} \quad (3)$$

IV. A simulation study

Simulation framework

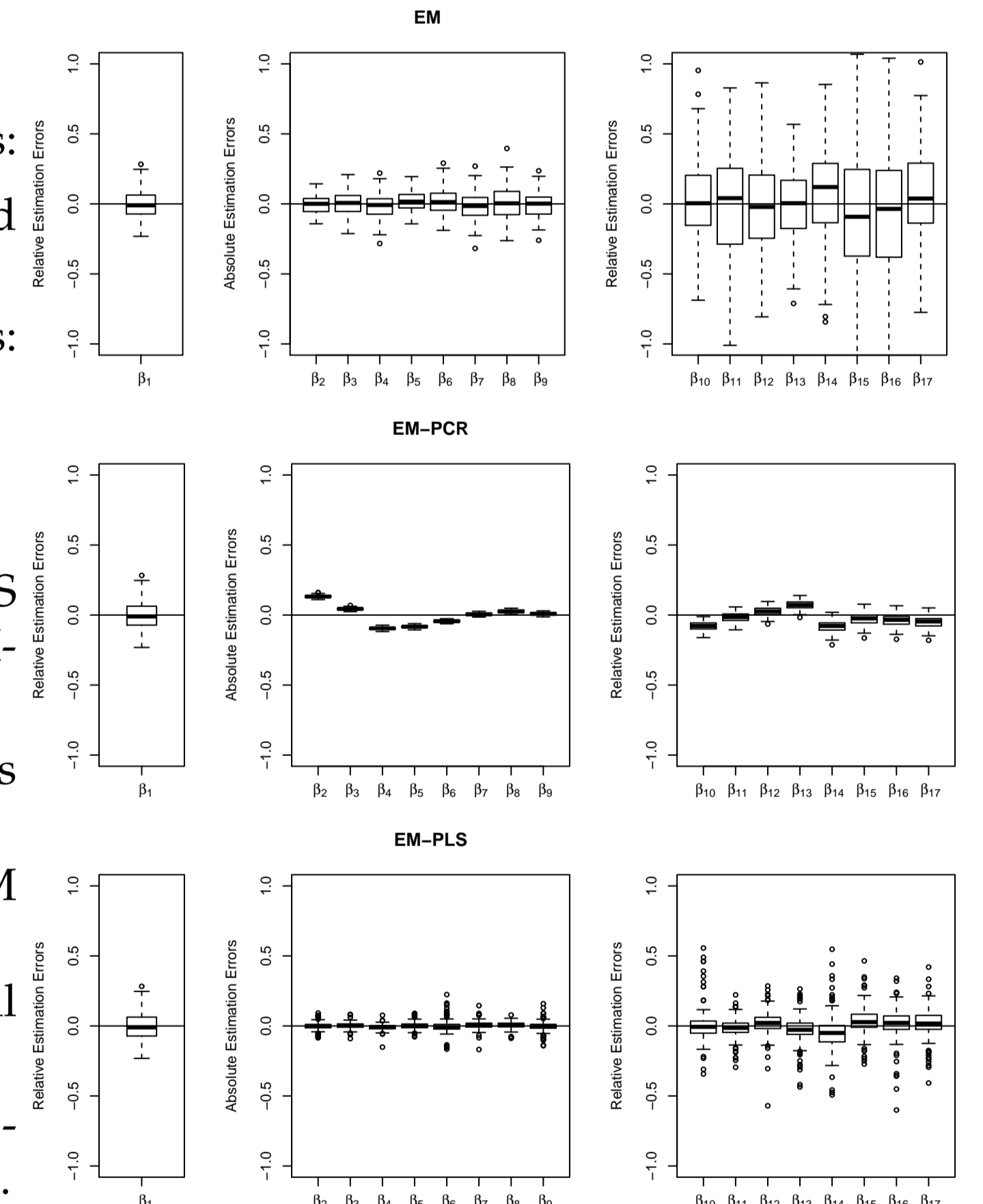
- For each individual i ($i = 1, \dots, 20$), $Y_i = X_i \beta + U_i \xi_i + \varepsilon_i$
- $X_i = (\mathbb{1}_{n_i}, X_i^1, X_i^2, X_i^3, X_i^4)$ where X_i^k is the $n_i \times 4$ fixed centered design matrix, $k = 1, \dots, 4$ and $U_i = \mathbb{1}_{n_i}$ with $n_i = 12 \quad \forall i = 1, \dots, 20$.
- $\beta = \{2.5, \{0\}^4, \{0\}^4, \{0.5\}^4\}$
- $\xi_i \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_i \sim \mathcal{N}(0_{12}, \sigma^2 I_{12})$ where σ^2 and τ^2 are respectively defined from a given signal-to-noise ratio SNR and a given variances ratio TAU by $\sigma^2 = \frac{1}{SNR^2} \frac{\|X\beta - E(X\beta)\|_2}{n}$ and $\tau^2 = \frac{\sigma^2}{TAU}$.
- $N = 100$ simulated data sets of size $n = 20 \times 12 = 240$ with different SNR and TAU values (learning set= 100 and test set=100).
- Criteria comparison:
 - Relative parameter estimation errors: $\frac{(\hat{\beta}_{jk} - \beta_j)}{\beta_j}$, $k = 1, \dots, 100$, $j = 1$ and $10, \dots, 17$
 - Absolute parameter estimation errors: $\hat{\beta}_{jk} - \beta_j$, $k = 1, \dots, 100$, $j = 2, \dots, 9$

Results for $SNR = 3$ and $TAU = 1$

→ Concerning the β_j 's estimation, EM-PLS method performs better than EM and EM-PCR methods:

- On average, good parameter estimations are obtained with EM-PLS method
- Large variability is obtained with EM method for all β_j different to 0
- EM-PCR method performs poorly for all β_j equal to 0

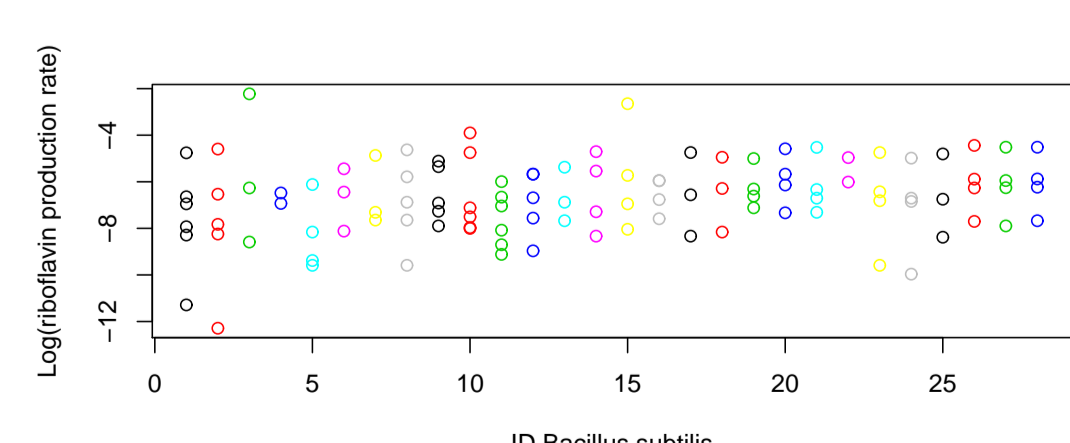
→ Globally, good estimation results are obtained with the three methods for τ^2 and σ^2 .



V. An application: the Riboflavin data set

- Data about riboflavin (vitamin B2) production in *Bacillus subtilis*.
 - Response variable : logarithm of the riboflavin production rate
 - Design matrix : the logarithm of the expressions levels of 4088 genes (normalized)
- 28 *Bacillus subtilis* with a total number of samples equal to 111
 - Observations at different times in the same conditions
 - From 2 to 6 measurements by *Bacillus subtilis*

Logarithm of the riboflavin production rate (vitamin B2) produced in 28 *Bacillus subtilis*. Different colors are used for the different *Bacillus subtilis*.

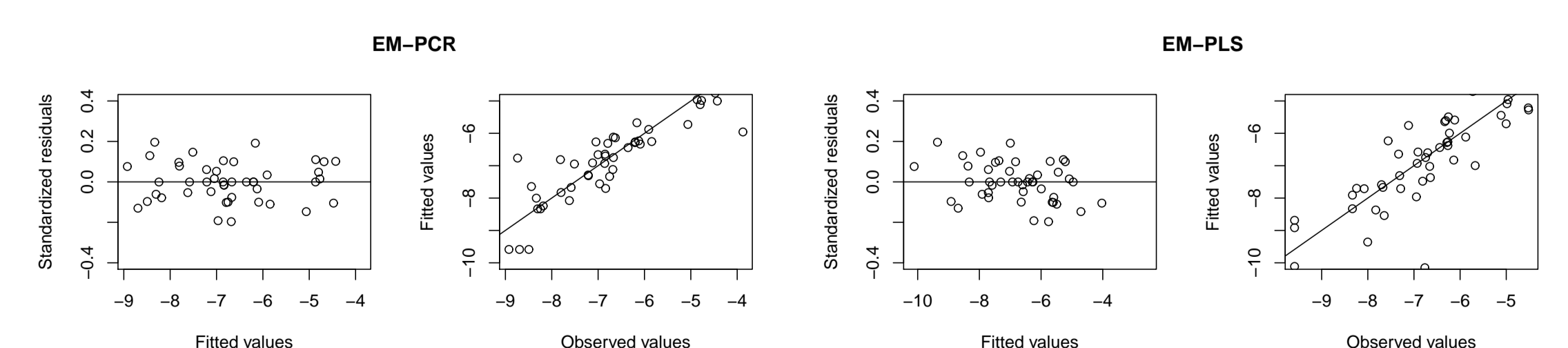


Method

- Subdivision of the data set into a learning set and a test set (ratio: 70 % - 30 %)
- Application of EM, EM-PCR and EM-PLS methods
- Computation of the mean absolute prediction error: $MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$
- Diagnostics plots: Normalized residuals vs fitted values and fitted values vs observed values plots.

Results

	EM-PCR	EM-PLS
Number of optimal components	8	8
MAE	0.43	0.63
$\hat{\sigma}^2$	0.62	0.02
$\hat{\tau}^2$	44.53	47.83



- No results with EM method because of numerical problems.
- Similar results are obtained with EM-PLS and EM-PCR methods.
- Possible improvement: a pre-selection step such as a Sure Independence Screening (SIS) procedure could be applied.

References

- Massy, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Stat. Assoc.*, 60:234-246.
- Helland, I. (1988). On the structure of Partial Least Squares Regression. *Commun. Stat., Simulation Comput.*, 17(2):581-607.

- Pinheiro, J., Bates, D. (2000). Mixed-effects models in S and S-Plus. Springer-Verlag, New-York.
- Lee, J.-M., Zhang, S., Saha, S., S.S. Anna, Jiang, C., and Perkins, J. (2001). RNA expression analysis using an antisense *Bacillus subtilis* genome array. *J. Bacteriol.*, 183:7371-7380.