Selection consistency in non-linear mixed-effects models under spike-and-slab prior

Marion Naveau, Laure Sansonnet, Maud Delattre

Selection consistency in non-linear mixed-effects models under spike-and-slab prior

Marion Naveau[1,2]
Direction : Maud Delattre[2], Laure Sansonnet[1]

[1]Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay

[2]Université Paris-Saclay, INRAE, MaIAGE

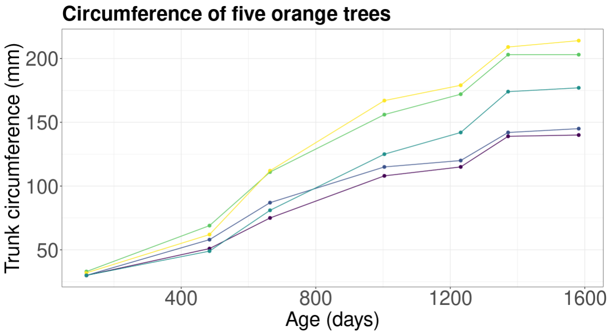
Colloque Jeunes Probabilistes et Statisticiens
2 Octobre 2023

Table of contents

1. Introduction
   - Framework
   - Prior specification
2. Theoretical guarantees
   - State of the art
   - First result
3. Perspectives

Framework: repeated measurement data

✤ **Mixed-effects models:** analyse observations collected repeatedly on several individuals.



✤ Same overall behaviour but with individual variations.
✤ Non-linear growth.
✤ Are these variations due to known characteristics?
- ▶ E.g.: growing conditions, genetic markers, ...

Non-linear mixed-effects model (NLMEM)

1) Description of intra-individual variability:
For all $i \in \{1, \dots, n\}$, $j \in \{1, \dots, n_i\}$,

$$y_{ij} = g(\varphi_i, \psi, t_{ij}) + \varepsilon_{ij},\ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

- $y_{ij} \in \mathbb{R}$: response of individual i at time t_{ij} (observation).
- $\varphi_i \in \mathbb{R}^q$: individual parameter, **not observed**.
- $\psi \in \mathbb{R}^r$: fixed effects, **unknown**.
- g: non-linear function with respect to φ_i (known).

2) Description of inter-individual variability:

$$\varphi_i = \mu + V_i\beta + \xi_i,\ \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, \Gamma)$$

- $\mu \in \mathbb{R}^q$: intercept, **unknown**.
- $V_i \in \mathbb{R}^p$: covariates for individual i (known).
- $\beta \in \mathcal{M}_{p\times q}$ covariate fixed effects matrix, **unknown**.

Population parameters: $\theta = (\mu, \beta, \psi, \sigma^2, \Gamma)$

High-dimensional covariate selection in NLMEM

✤ Goal: identify

$$S_0 = \left\{ (\ell, m) \in \{1, \ldots, p\} \times \{1, \ldots, q\} \middle| \beta^0_{\ell m} \neq 0 \right\},$$

where β^0 is the true fixed effects matrix.

✤ Specificity of the problem: $p >> n$

$\Rightarrow$ Assumption: each column of β^0 is sparse

✤ Main difficulties:

- High-dimensional variable selection:
  - ▶ parsimonious estimation of β
    - ➢ regularised methods (LASSO-type, Tibshirani (1996))
    - ➢ sparsity-inducing priors (Tadesse and Vannucci, 2021)
- Non-explicit likelihood
  - ▶ The φ_i's are not observed (latent variables model)
    - ➢ theoretical and algorithmic in LMEM (Schelldorfer et al., 2011)
  - ▶ g is non-linear
    - ➢ algorithmic only in NLMEM (Ollier (2021), not high-dimension)

Methodology

Proposed approach

Association of a Bayesian ***spike-and-slab*** Gaussian prior for variable selection with a stochastic version of the EM algorithm, called **MCMC-SAEM**, for inference.

Naveau, M., Kon Kam King, G., Rincent, R., Sansonnet, L., and Delattre, M. (2022). **Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm.** arXiv preprint arXiv:2206.01012.

Spike-and-slab prior for the coefficients of β

✤ Introduction of latent variables $\delta_{\ell m}$, $1 \leq \ell \leq p$, $1 \leq m \leq q$:

$$\delta_{\ell m} = \begin{cases} 1 & \text{if } (\ell, m) \text{ is to be included in model } S^*, \\ 0 & \text{otherwise.} \end{cases}$$

✤ **Spike-and-slab prior** on β (George and McCulloch, 1997):

$$\pi(\beta_{\ell m}|\delta_{\ell m}) = (1 - \delta_{\ell m})h_0(\beta_{\ell m}) + \delta_{\ell m}h_1(\beta_{\ell m}),$$

i.e.:

- $\beta_{\ell m}|(\delta_{\ell m} = 0) \sim h_0(\beta_{\ell m})$: concentrated "spike" distribution
- $\beta_{\ell m}|(\delta_{\ell m} = 1) \sim h_1(\beta_{\ell m})$: diffuse "slab" distribution

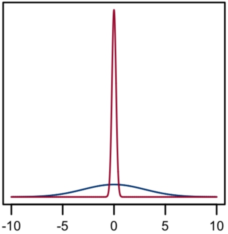


Figure: Spike-and-slab Gaussian prior. Source: Deshpande et al. (2019)

1. Introduction
   - Framework
   - Prior specification
2. Theoretical guarantees
   - State of the art
   - First result
3. Perspectives

Selection consistency

✤ Aim: obtain selection consistency results for non-linear mixed effects models under spike-and-slab prior.

✤ Selection consistency: the posterior probability of the true model converges to 1, $\inf_{\beta_0} \mathbb{E}_0 \left[\Pi(\beta : S_\beta = S_0 | Y^{(n)}) \right] \longrightarrow 1$.

✤ Posterior contraction: ability of the posterior distribution to recover the true model from the data
$\sup_{\theta_0} \mathbb{E}_0 \left[\Pi \left(\theta : d_n(\theta, \theta_0) > Cn\epsilon_n \middle| Y^{(n)} \right) \right] \longrightarrow 0$, with $\epsilon_n \longrightarrow 0$.

State of the art

✤ **With known variance:**

Reference	Model	Spike	Slab	Result
Castillo et al. (2015)	LR	Dirac	Laplace	Consistency
Ročková and George (2018)	LR	Laplace	Laplace	Contraction

✤ **With unknown variance:**

Reference	Model	Spike	Slab	Result
Narisetty and He (2014)	LR	Gaussian	Gaussian	Consistency
Jiang and Sun (2019)	LR	Generic	Generic	Consistency
Ning et al. (2020)	Multivariate LR	Dirac	Laplace	Consistency
Jeong and Ghosal (2021a)	GLMs	Dirac	Laplace	Contraction
Jeong and Ghosal (2021b)	LR with nuisance	Dirac	Laplace	Consistency
Shen and Deshpande (2022)	Multivariate LR	Laplace	Laplace	Contraction

where LR = Linear Regression.

Stages of proof

In general, the stages of proof (following Castillo et al. (2015)) are as follows:

1. **Support size:** $\sup_{\beta_0} \mathbb{E}_0 \left[\Pi \left(\beta : |S_\beta| > K|S_0| \middle| Y^{(n)} \right) \right] \longrightarrow 0$
2. **Posterior contraction / Recovery:**
$$\sup_{\theta_0} \mathbb{E}_0 \left[\Pi \left(\theta : d_n(\theta, \theta_0) > C n \epsilon_n \middle| Y^{(n)} \right) \right] \longrightarrow 0, \text{ with } \epsilon_n \longrightarrow 0$$
3. **Distributional approximation:**
$$\sup_{\beta_0} \mathbb{E}_0 \left[\left\| \Pi \left(\beta \in \cdot | Y^{(n)} \right) - \Pi^\infty \left(\beta \in \cdot | Y^{(n)} \right) \right\|_{TV} \right] \longrightarrow 0$$
4. **Selection, no supersets:** $\sup_{\beta_0} \mathbb{E}_0 \left[\Pi \left(\beta : S_\beta \supset S_0, S_\beta \neq S_0 \middle| Y^{(n)} \right) \right] \longrightarrow 0$
5. **Selection consistency:** $\inf_{\beta_0} \mathbb{E}_0 \left[\Pi(\beta : S_\beta = S_0 | Y^{(n)}) \right] \longrightarrow 1.$

First approach: extension of Jeong and Ghosal (2021b)

- Consider a non-linear marginal mixed model:

$$y_i = f_i(\beta) + Z_i\xi_i + \varepsilon_i^*,\ \varepsilon_i^* \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i}),\ \xi_i \sim \mathcal{N}_q(0, \Gamma),$$

where $y_i \in \mathbb{R}^{n_i}$, $\beta \in \mathbb{R}^p$, $Z_i \in \mathcal{M}_{n_i \times q}$ and $n_i \in \{1, \dots, J_n\}$, where n_i and J_n can grow with n. We assumed that σ^2 is known.

- Thus, this model can be re-written as:

$$y_i = f_i(\beta) + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I + Z_i \Gamma Z_i^\top),$$

We denote by S the support of β. The true parameters are noted: β_0, Γ_0, S_0 and $s_0 = |S_0|$.

- Prior:
  - Spike-and-slab Dirac-Laplace on (S, β): $(S, \beta) \mapsto \frac{\pi_p(s)}{\binom{p}{s}} g_S(\beta_S)\delta_0(\beta_{S^c})$,
  - Inverse-Wishart(Σ, d) prior on Γ: $\pi(\Gamma) \propto |\Gamma|^{-(d+q+1)/2} \exp\left(-\frac{1}{2} Tr(\Sigma\Gamma^{-1})\right)$.

Assumptions

- For each i, f_i is assumed to be Lipschitzienne:

$$||f_i(\beta) - f_i(\tilde{\beta})||_2 \leq K_i||\beta - \tilde{\beta}||_2, \text{ for all } \beta, \tilde{\beta} \in \mathbb{R}^p.$$

  We denote by $K = \sum_{i=1}^n K_i^2$.

  ▶ **Example:** Log-Gompertz model $y_{ij} = \beta_1 + b_i - Ce^{-\beta_2 t_{ij}} + \varepsilon_{ij}$

- $g_S(\beta_S) = \prod_{j \in S} \frac{\lambda}{2} \exp(-\lambda|\beta_j|)$, with $\frac{\sqrt{K}}{L_1 p^{L_2}} \leq \lambda \leq \frac{L_3\sqrt{K}}{\sqrt{n}}$, for some constants L_1, L_2, $L_3 > 0$.

- For some constants A_1, A_2, A_3, $A_4 > 0$,

$$A_1 p^{-A_3}\pi_p(s-1) \leq \pi_p(s) \leq A_2 p^{-A_4}\pi_p(s-1),\ s = 1, \dots p.$$

  ▶ **Example:** $\beta_1, \dots \beta_p \sim (1-r)\delta_0 + r\mathcal{L}$, $\pi_p = Bin(p, r)$ where $r \sim Beta(1, p^u)$, $u > 1$.

Assumptions

- $s_0 > 0$,
- $s_0 \log(p) = o(n)$, $\frac{\log(n)}{\log(p)} \to 0$, $\log(J_n) \lesssim \log(p)$
- $1 \lesssim \rho_{min}(\Gamma_0) \leq \rho_{max}(\Gamma_0) \lesssim 1$,
- $||\beta_0||_\infty \lesssim \lambda^{-1} \log(p)$,
- $\dfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{n_i \geq q}$ is bounded,
- $\min_i \{\rho_{\min}^{1/2}(Z_i^\top Z_i) : n_i \geq q\} \gtrsim 1$, *i.e.* Z_i is a full rank,
- $\max_i \{\rho_{\max}^{1/2}(Z_i^\top Z_i)\} \lesssim 1$.

Support size theorem

Theorem

Let's assume that the previous hypotheses are satisfied. Then, there exists a constant C_1 such that:

$$\mathbb{E}_0\left[\Pi\left(\beta : |S_\beta| > C_1 s_0 \middle| y\right)\right] \longrightarrow 0.$$

1. Introduction
   - Framework
   - Prior specification
2. Theoretical guarantees
   - State of the art
   - First result
3. Perspectives

Perspectives

✤ In non-linear: Under spike-and-slab Dirac-Laplace, can we get:

- Posterior contraction?
- Distributional approximation of the posterior?
- Selection consistency?

under what assumptions?
Can the same results be obtained by making the model more complex (Z_i depending on β)?

✤ In linear: **Can we obtain a selection consistency theorem under spike-and-slab LASSO prior in LMEM with covariance matrix unknown?**

Thank you for your attention!

References I

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018. Publisher: Institute of Mathematical Statistics.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

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

Deshpande, S. K., Ročková, V., and George, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, 28(4):921–931.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

References II

Jeong, S. and Ghosal, S. (2021a). Posterior contraction in sparse generalized linear models. *Biometrika*, 108(2):367–379. Publisher: Oxford University Press.

Jeong, S. and Ghosal, S. (2021b). Unified Bayesian theory of sparse linear regression with nuisance parameters. *Electronic Journal of Statistics*, 15(1):3040–3111. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Jiang, B. and Sun, Q. (2019). Bayesian high-dimensional linear regression with generic spike-and-slab priors. *arXiv preprint arXiv:1912.08993*.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817. Publisher: Institute of Mathematical Statistics.

Ning, B., Jeong, S., and Ghosal, S. (2020). Bayesian linear regression for multivariate responses under group sparsity.

References III

Ollier, E. (2021). Fast selection of nonlinear mixed effect models using penalized likelihood. *arXiv preprint arXiv:2103.01621*.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444. Publisher: Taylor & Francis.

Schelldorfer, J., Bühlmann, P., and DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

Shen, Y. and Deshpande, S. K. (2022). On the posterior contraction of the multivariate spike-and-slab LASSO. *arXiv preprint arXiv:2209.04389*.

Tadesse, M. G. and Vannucci, M. (2021). *Handbook of Bayesian variable selection*. CRC Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Model approximation

$$\begin{cases} y_i = f_i(\psi, \varphi_i) + \varepsilon_i & , \varepsilon_i \overset{\text{ind}}{\sim} \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i}), \\ \varphi_i = X_i\beta + \xi_i & , \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, \Gamma). \end{cases}$$

where $y_i \in \mathbb{R}^{n_i}$, $f_i(\psi, \varphi_i) = (f(\psi, \varphi_i; t_{i,1}), \ldots, f(\psi, \varphi_i; t_{i,n_i}))$, $\psi \in \mathbb{R}^r$, $\varphi_i \in \mathbb{R}^q$, $X_i \in \mathcal{M}_{q\times p}$, $\beta \in \mathbb{R}^p$.

First order approximation of $f_i(\psi, X_i\beta + \xi_i)$ around $\mathbb{E}[\varphi_i] = X_i\beta$:

$$y_i = f_i(\psi, X_i\beta) + Z_i(\beta)\xi_i + \varepsilon_i,$$

where $Z_i = \dfrac{\partial f_i}{\partial \varphi_i}$.

$\Longrightarrow$ Non-linear marginal mixed model with varied matrix of random effects (Demidenko, 2013).

Idea of the proof

Set $B = \{(\beta, \Gamma) : |S_\beta| > \tilde{s}\}$, with any integer $\tilde{s} \geq s_0$.

Yet, by Bayes' formula: $\Pi(B|y) = \dfrac{\int_B \Lambda_n(\beta,\Gamma) d\Pi(\beta,\Gamma)}{\int \Lambda_n(\beta,\Gamma) d\Pi(\beta,\Gamma)}$, where

$\Lambda_n(\beta,\Gamma) = \prod_{i=1}^n \dfrac{p_{\beta,\Gamma,i}}{p_{0,i}}$ likelihood ratio.

Thus, the following lemma shows that the denominator of the posterior distribution is bounded below by a factor with probability tending to one:

Lemma

Let's assume that the previous hypotheses are satisfied. Then, there exists a constant M such that:

$$\mathbb{P}_0 \left(\int \Lambda_n(\beta,\Gamma) d\Pi(\beta,\Gamma) \geq \pi_p(s_0) e^{-M(s_0 \log(p) + \log(n))} \right) \longrightarrow 1.$$

This event is denoted by $\mathcal{A}_n$.

Idea of the proof

Then, $\mathbb{E}_0\left[\Pi\left(B|y\right)\right] = \mathbb{E}_0\left[\Pi\left(B|y\right)\mathbb{1}_{\mathcal{A}_n}\right] + \underbrace{\mathbb{E}_0\left[\Pi\left(B|y\right)\mathbb{1}_{\mathcal{A}_n^c}\right]}_{\longrightarrow 0 \text{ by lemma}}$.

And by the lemma and Fubini-Tonelli's theorem the first term is bounded by a term tending towards 0 with n:

$$\begin{aligned}\mathbb{E}_0\left[\Pi\left(B|y\right)\mathbb{1}_{\mathcal{A}_n}\right] &= \mathbb{E}_0\left[\frac{\int_B \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}{\int \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}\mathbb{1}_{\mathcal{A}_n}\right] \\ &\leq \pi_p(s_0)^{-1}\exp\left\{M(s_0\log(p)+\log(n))\right\}\Pi(B) \longrightarrow 0.\end{aligned}$$

This leads to the theorem: there exist a constant C_1 such that $\mathbb{E}_0\left[\Pi\left(|S_\beta| > C_1 s_0|y\right)\right] \longrightarrow 0$.

Posterior contraction Rényi theorem

Definition

For two n-variates densities $f = \prod_{i=1}^{n} f_i$ and $g = \prod_{i=1}^{n} g_i$ of independent variables, the average Rényi divergence (of order $1/2$) is defined by:

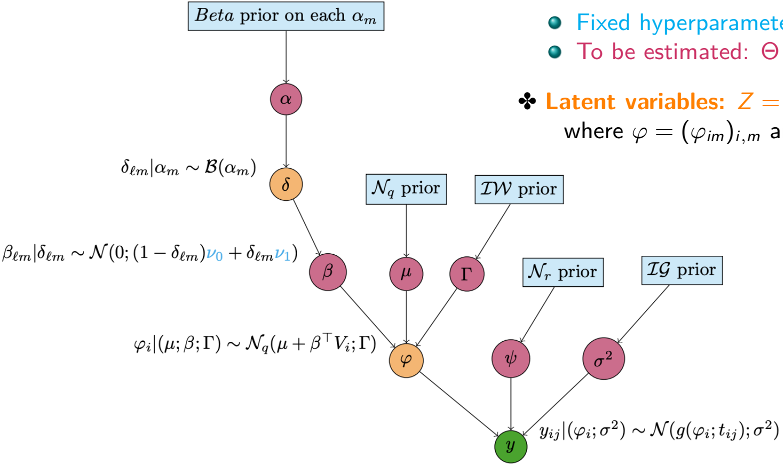
$$R_n(f, g) = -\frac{1}{n}\sum_{i=1}^{n} \log \int \sqrt{f_i g_i}$$

Theorem

Let's assume that the previous hypotheses are satisfied. We denote by $p_{\beta,\Gamma} = \prod_{i=1}^{n} p_{\beta,\Gamma,i}$ the joint density for $p_{\beta,\Gamma,i}$ the density of the ith observation vector y_i, and p_0 the true joint density. Then, there exists a constant C_2 such that:

$$\mathbb{E}_0\left[\Pi\left((\beta,\Gamma) : R_n(p_{\beta,\Gamma}, p_0) > C_2\frac{s_0\log(p)}{n}\middle| y\right)\right] \longrightarrow 0.$$

Bayesian hierarchical model



- **Observations:** $y = (y_{ij})_{i,j}$
- **Parameters:**
  - Fixed hyperparameters: ν_0, ν_1, ...
  - To be estimated: $\Theta = (\theta, \alpha)$
- **Latent variables:** $Z = (\varphi, \delta)$
  where $\varphi = (\varphi_{im})_{i,m}$ and $\delta = (\delta_{\ell m})_{\ell,m}$

Proposed method

Idea: explore different levels of sparsity in each column of β by varying the value of ν_0 in a grid Δ.

1. **Creation of a model collection:** for each $\nu_0 \in \Delta$,
   - Compute $\widehat{\Theta}$ by a MCMC-SAEM algorithm (Kuhn and Lavielle, 2004):
   $$\widehat{\Theta}_{\nu_0}^{MAP} = \underset{\Theta \in \Lambda}{\text{argmax}}\ \pi(\Theta|y)$$
   - Estimate $\hat{\delta}$ (Ročková and George, 2014): $\hat{\delta} = \underset{\delta}{\text{argmax}}\ P(\delta|\hat{\Theta}_{\nu_0}^{MAP})$ such as

   $\hat{\delta}_{\ell m} = 1 \iff \mathbb{P}(\delta_{\ell m} = 1|\hat{\Theta}_{\nu_0}^{MAP}) \geq 0.5 \iff$ Define:
   $$\widehat{S}_{\nu_0} = \left\{ (\ell, m) \in \{1, \ldots, p\} \times \{1, \ldots, q\} \;\middle|\; |(\widehat{\beta}_{\nu_0}^{MAP})_{\ell m}| \geq s_\beta(\nu_0, \nu_1, (\widehat{\alpha}_{\nu_0}^{MAP})_m) \right\}$$

2. **Select the "best" model** among $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$ by a fast criterion, eBIC (Chen and Chen, 2008):
   $$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\text{argmin}} \left\{ -2 \log\left(p(y; \hat{\theta}_{\nu_0}^{MLE})\right) + B_{\nu_0} \times \log(n) + 2 \log\left(\binom{pq}{B_{\nu_0}}\right) \right\}$$
   with B_{ν_0}: number of free parameters in the model $\widehat{S}_{\nu_0}$.

3. **Return** $\widehat{S}_{\hat{\nu}_0}$.

Spike-and-slab regularisation plot

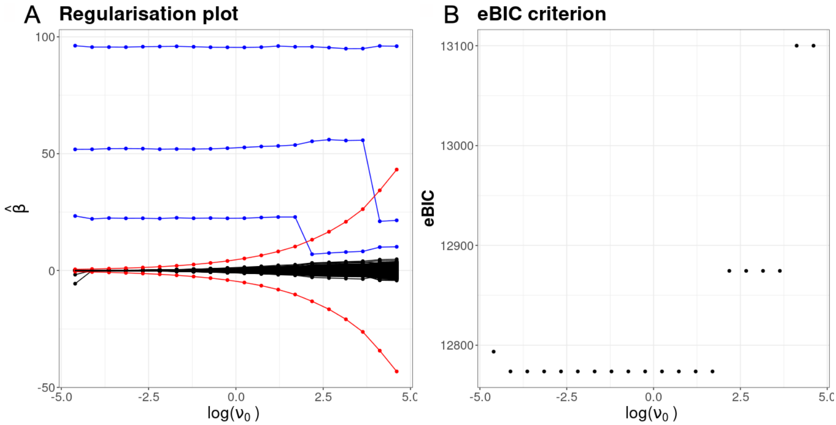


Figure: $q = 1,\ n = 200,\ J = 10,\ p = 500,\ \gamma^2 = 200,\ \sigma^2 = 30,\ \nu_1 = 12000,$ $\mu = 1200,\ \beta = {}^t(100, 50, 20, 0, \ldots, 0)$

Computing the MAP in a latent variables model

✤ Let's go back to the **first step** of the proposed method:

- Compute the MAP estimator of Θ
- Goal: maximise $\pi(\Theta|y) = \int_{\mathcal{Z}} \pi(\Theta, Z|y) dZ$ with

$$\pi(\Theta, Z|y) = \frac{p(y|\Theta, Z)p(\Theta, Z)}{\int_{\mathcal{Z}} \int_{\Lambda} p(y|\Theta, Z)p(\Theta, Z) d\Theta dZ}$$

- Non-explicit integral

EM algorithm (Dempster et al., 1977)

1. Initialisation: choose $\Theta^{(0)}$.
2. Iteration $k \geq 0$:
    - **E-step (Expectation):** compute

$$Q(\Theta|\Theta^{(k)}) = \mathbb{E}_{Z|(y,\Theta^{(k)})}\left[\log(\pi(\Theta, Z|y))\middle| y, \Theta^{(k)}\right].$$

    - **M-step (Maximisation):** compute

$$\Theta^{(k+1)} = \underset{\Theta\in\Lambda}{\operatorname{argmax}}\; Q(\Theta|\Theta^{(k)}).$$

3. $\hat{\Theta} = \Theta^{(K)}$, for K large enough.