Compte rendu de travail pour le 28/10/2022

1- Ce que j'ai fait cette semaine

1. En section 2, la formulation du calcul de Q par récurrence ainsi qu'un petit paragraphe sur sa complexité.
2. En section 3 et 4 j'ai travaillé sur Dempster, Laird, and Rubin 1977 en particulier la démonstration de convergence, Je n'ai fait que reprendre, reformuler pour pouvoir l'assimiler. J'avais en particulier besoin de comprendre ce qu'était un GEM. **Vous pouvez omettre cette partie**
3. En section 5, j'ai détaillé l'intégration du proximal et de la pénalisation en reprenant Fort, Ollier, and Samson 2017.
4. En section 6, j'ai voulu par curiosité voir à quoi ressemblerait le modèle joint inspiré par He et al. 2015. J'ai également écrit une question de modélisation que je n'ai pas réussi à répondre seul !
5. En annexe 6, j'ai remis l'algorithme corrigé du stage que je vous avez envoyé il y 2 semaines, Si jamais vous voulez comparer avec l'algo de la section 5. Mais ce sont pour moi les mêmes !

pvi : J'ai crée un répertoir git de partage : Stat4Plant_WP3

2- SAEM without exponential family

We wish to write here the approximation step in a general context according to the likelihood only.

Assuming that the A step of the SAEM in the general context is written :

$$Q_{k+1}(\theta) = (1-u_{k+1})Q_k(\theta) + u_{k+1}\ell_{comp}(\theta; Z^{(k+1)}; Y)$$

where $\ell_{comp} = \log \mathcal{L}_{comp}$, Z is some latent variable and Y the observation.

Proposition

$$Q_{k+1}(\theta) = \left(\prod_{l=1}^{k+1}(1-u_l)\right) Q_0(\theta) + \sum_{l=1}^{k+1}\left(\prod_{h=l}^{k}(1-u_{h+1})\right) \times u_l \ell_{comp}(\theta; Z^{(l)}; Y)$$

Proof

Let us define the recurrence proposal according to our proposition. For $k = 0$, we have obviously that : $Q_1 = (1-u_1)Q_0(\theta) + u_1\ell_{comp}(\theta; Z^{(1)}; Y)$ so $\mathcal{P}_0$ true !

Let be for a certain $k \geqslant 0$, suppose that $\mathcal{P}_k$ is true, we have :

$$\begin{aligned}
Q_{k+2} =& (1-u_{k+2})Q_{k+1} + u_{k+2}\ell_{comp}(\theta; Z^{(k+2)}; Y) \\
=& (1-u_{k+2})\left(\prod_{l=1}^{k+1}(1-u_l)\right) Q_0(\theta) \\
&+ \sum_{l=1}^{k+1}\left(\prod_{h=l}^{k}(1-u_{h+1})\right) \times u_l \ell_{comp}(\theta; Z^{(l)}; Y) \\
&+ u_{k+2}\ell_{comp}(\theta; Z^{(k+2)}; Y) \\
=& \left(\prod_{l=1}^{k+2}(1-u_l)\right) Q_0(\theta) + \sum_{l=1}^{k+1}\left(\prod_{h=l}^{k+1}(1-u_{h+1})\right) \times u_l \ell_{comp}(\theta; Z^{(l)}; Y) \\
&+ u_{k+2}\ell_{comp}(\theta; Z^{(k+2)}; Y) \\
=& \left(\prod_{l=1}^{k+2}(1-u_l)\right) Q_0(\theta) + \sum_{l=1}^{k+2}\left(\prod_{h=l}^{k+1}(1-u_{h+1})\right) \times u_l \ell_{comp}(\theta; Z^{(l)}; Y)
\end{aligned}$$

QED □

So if we initiate Q with $Q_0 = 0$, we have $Q_{k+1}(\theta) = \sum_{l=1}^{k+1}\left(\prod_{h=l}^{k}(1-u_{h+1})\right) \times u_l \ell_{comp}(\theta; Z^{(l)}; Y)$.

Suppose we want to run a SAEM for a mixed effect model with $N \times J$ observations where N is the number of individuals and J is the number of observations for each individual. Suppose that the mixed effect variables are of dimension N and that the maximization step is in a constant time $O(1)$. Let's say that the computation of ℓ_{comp} **takes a constant time** and depend only on the total number of observations so computation will be in $O(NJ)$. Achieve the approximation step, i.e., the computation of Q_k, will take about a linear time $O(k \times NJ)$. To compare, the calculation of the simulation step, following a one-step MCMC procedure, is computed in $O(N \times J)$ if it has been well coded and in the more casual case, i.e., without optimization procedures, in $O(N \times NJ)$.

To conclude if we do a total of K iteration of the SAEM, the complexity of all the approximation step will be $O(K^2NJ)$. The simulation step is done in $O(KNJ)$. **This algorithm has a quadratic complexity in time depending on the number of iterations !**

3- EM algorithm definition

The goal of the algorithm is to find a local maximum likelihood of a model that depends on unobserved latent variables :

$$\arg\max_{\theta\in\mathbb{R}^p}\{\log \mathcal{L}_{marg}(Y,\theta)\}$$

The EM focus on maximizing the quantity $Q(\theta|\theta_k)$ rather than directly improve $\log \mathcal{L}_{marg}(Y;\theta)$ (Dempster, Laird, and Rubin 1977).

Algorithm 1: Expectation Maximization

Require : Number of iterations $K \geqslant 1$
Initialize $\theta_0 \in \mathbb{R}^d$ initial parameter value
for $k = 1$ **to** K **do**
- **Step E :** computation of $Q(\theta|\theta_k) = \mathbb{E}[\log \mathcal{L}_{comp}(Y,Z;\theta)|Y;\theta_k]$
- **Step M :** maximization of $Q(.|\theta_k) : \theta_{k+1} = \arg\max_{\theta\in\mathbb{R}^p} Q(\theta|\theta_k)$

end
return $\hat{\theta} = \theta_K$

4- General properties from Dempster, Laird, and Rubin 1977

We introduce the notation for the conditional density of Z given Y and θ : $p(Z|Y;\theta) = \frac{\mathcal{L}_{comp}(Y,Z;\theta)}{\int \mathcal{L}_{comp}(Y,Z;\theta)dZ} = \frac{\mathcal{L}_{comp}(Y,Z;\theta)}{\mathcal{L}_{marg}(Y;\theta)}$. As $\log p(Z|Y;\theta) = \log \mathcal{L}_{comp}(Y,Z;\theta) - \log \mathcal{L}_{marg}(Y;\theta)$, the log marginal likelihood can be written :

$$\boxed{\log \mathcal{L}_{marg}(Y;\theta) = \log \mathcal{L}_{comp}(Y,Z;\theta) - \log p(Z|Y;\theta)}$$

We take the expectation value of the observed data Z given an estimate of the parameter θ'. To do this we multiply by the density of Z and integrate over Z :

$$\begin{aligned}\log \mathcal{L}_{marg}(Y;\theta) =& \int \log \mathcal{L}_{comp}(Y,Z;\theta)p(Z|Y,\theta')dz - \int \log p(Z|Y;\theta)p(Z|Y,\theta')dz \\ =& \mathbb{E}[\log \mathcal{L}_{comp}(Y,Z;\theta)|Y,\theta'] - \mathbb{E}[\log p(Z|Y;\theta)|Y,\theta']\end{aligned}$$

We use the already established notation : $\boxed{\begin{cases} Q(\theta|\theta') &= \mathbb{E}[\log \mathcal{L}_{comp}(Y,Z;\theta)|Y,\theta'] \\ H(\theta|\theta') &= \mathbb{E}[\log p(Z|Y;\theta)|Y,\theta'] \end{cases}}$

$$\log \mathcal{L}_{marg}(Y;\theta) = Q(\theta|\theta') - H(\theta|\theta')$$

Lemma *Lemma 1 from Dempster, Laird, and Rubin 1977*

$$H(\theta',\theta) \leqslant H(\theta|\theta), \forall(\theta',\theta) \in \mathbb{R}^p \times \mathbb{R}^p$$

Je n'ai pas réussi / approfondie la démo de ce lemme. Il se fait apparement avec des formules de Jensen provenant d'un ouvrage que je n'ai pas trouvé en libre accès.

Definition: *Generalized-EM algorithm (GEM)*

Let M be a mapping : $\theta \mapsto M(\theta)$ from $\mathbb{R}^p$ to $\mathbb{R}^p$ such that each step $\theta_k \to \theta_{k+1}$ of the EM is defined by :

$$\theta_{k+1} = M(\theta)$$

An iterative algorithm with mapping M is a **generalized-EM algorithm** if :

$$Q(M(\theta)|\theta) \geqslant Q(\theta|\theta), \forall \theta \in \mathbb{R}^p$$

Theorem *Theorem 1 of Dempster, Laird, and Rubin 1977*

For every GEM algorithm, we have :

$$\log \mathcal{L}_{marg}(Y; M(\theta)) \geqslant \log \mathcal{L}_{marg}(Y; \theta), \forall \theta \in \mathbb{R}^p$$

5- Introduction of the proximal operator in the EM

We want to introduce a penalty term in the likelihood and have the following optimization problem:

$$\arg\max_{\theta \in \mathbb{R}^p} \{\log \mathcal{L}_{marg}(Y, \theta) - \text{pen}(\theta)\}$$

We want to use the EM algorithm but with the penalty term, as :

Algorithm 2: Expectation Maximization - penalized

Require : Number of iterations $K \geqslant 1$

Initialize $\theta_0 \in \mathbb{R}^d$ initial parameter value

for $k = 1$ **to** K **do**

- **Step E :**
  computation of $Q(\theta|\theta_k) = \mathbb{E}[\log \mathcal{L}_{comp}(Y, Z; \theta)|Y; \theta_k]$
- **Step M :**
  maximization of $Q(.|\theta_k)$: $\theta_{k+1} = \arg\max_{\theta \in \mathbb{R}^p} \{Q(\theta|\theta_k) - pen(\theta)\}$

end

return $\hat{\theta} = \theta_K$

Fort, Ollier, and Samson proposes the following algorithm to obtain a maximum likelihood. **The presence of the proximal operator is explained in the following** (see 1).

Algorithm 3: proximal Expectation Maximization - penalized

Require : Number of iterations $K \geqslant 1$; a sequence of steps $\gamma_k > 0$

Initialize $\theta_0 \in \mathbb{R}^d$ initial parameter value

for $k = 1$ **to** K **do**

- **Step E :**
  computation of $Q(\theta|\theta_k) = \mathbb{E}[\log \mathcal{L}_{comp}(Y, Z; \theta)|Y; \theta_k]$
- **Step M :**
  proximal-gradient descent : $\theta_{k+1} = \underset{\gamma,\text{pen}}{\text{Prox}}\{\theta_k + \gamma_{n+1}\nabla_1 Q(\theta|\theta_k)\}$

end

return $\hat{\theta} = \theta_K$

where ∇_1 denotes the gradient operator according to the first variable, and the Prox operator is defined as :

Definition: *Proximal operator*

Let γ be a positive step size,

$$prox_{\gamma,\text{pen}}(\theta) = \underset{\theta' \in \mathbb{R}^p}{\arg\min} \left(\text{pen}(\theta') + \frac{1}{2\gamma} \|\theta - \theta'\|_2^2 \right)$$

Definition: *Generalized-EM-penalized algorithm (GEM-pen)*

Let M be a mapping : $\theta \mapsto M(\theta)$ from $\mathbb{R}^p$ to $\mathbb{R}^p$ such that each step $\theta_k \to \theta_{k+1}$ of the EM is defined by :

$$\theta_{k+1} = M(\theta)$$

An iterative algorithm with mapping M is a **generalized-EM-pen algorithm** if :

$$Q(M(\theta)|\theta) - \text{pen}(M(\theta)) \geqslant Q(\theta|\theta) - \text{pen}(\theta), \forall \theta \in \mathbb{R}^p$$

Je ne sais pas si cette définition est original au papier d'Ollier

Proposition *Proposition 1 from Fort, Ollier, and Samson 2017*

Let assume that $\text{pen} : \mathbb{R}^p \to \mathbb{R}$ is convex and lower semi-continuous,
and that there exists a constant $L > 0$ such that for any $\theta' \in \mathbb{R}^p$ the gradient of $\theta \mapsto Q(\theta, \theta')$ is L-Lipschitz.
Let $(\gamma_k)_{k \geqslant 0}$ be a positive sequence of step-size, **such that** $\gamma_k \in]0, 1/L], \forall k \geqslant 0$.
Then the algorithm 3 is a GEM-algorithm for the maximization of $\log \mathcal{L}_{marg} - \text{pen}$

Proof

Lemma

Under the assumptions of 1, for any $\gamma \in]0, 1/L]$ and $\theta, \theta' \in \mathbb{R}^p \times \mathbb{R}^p$

$$Q(\theta|\theta') \geqslant Q(\theta|\theta') - \frac{1}{2\gamma}\left\|\gamma\nabla_1 Q(\theta', \theta') + \theta' - \theta\right\|^2 + \frac{\gamma}{2}\left\|\nabla_1 Q(\theta', \theta')\right\|^2$$

where ∇_1 denotes the gradient operator according to the first variable

Proof

Let $\theta, \theta' \in \mathbb{R}^p \times \mathbb{R}^p$,

Recall for any $\theta' \in \mathbb{R}^p$ the gradient of $Q(., \theta')$ is L-Lipschitz, so the Taylor's theorem give to ordre 1 at θ'

$$\begin{aligned} Q(\theta|\theta') &\geqslant Q(\theta'|\theta') + \left\langle \nabla_1 Q(\theta', \theta'); \theta - \theta' \right\rangle + \frac{1}{2}(\theta - \theta')^T \underbrace{\nabla_1^2 Q(\theta'|\theta')}_{\geqslant -L}(\theta - \theta') \\ &\geqslant Q(\theta'|\theta') + \left\langle \nabla_1 Q(\theta', \theta'); \theta - \theta' \right\rangle - \frac{L}{2}\left\|\theta - \theta'\right\|^2 \end{aligned}$$

We conclude using the inequality : $2\langle b, a\rangle - \|a\|^2 = \|b\|^2 - \|a - b\|^2$ with $a = \sqrt{L}(\theta - \theta')$ and $b = \frac{1}{\sqrt{L}}\nabla Q(\theta'|\theta')$

$$Q(\theta|\theta') \geqslant Q(\theta'|\theta') + \frac{1}{2L}\left\|\nabla_1 Q(\theta'|\theta')\right\|^2 - \frac{L}{2}\left\|\frac{1}{L}\nabla_1 Q(\theta'|\theta') + \theta - \theta'\right\|^2$$

Then with $\frac{1}{\gamma} \geqslant L$: $\boxed{Q(\theta|\theta') \geqslant Q(\theta'|\theta') + \frac{\gamma}{2}\left\|\nabla_1 Q(\theta'|\theta')\right\|^2 - \frac{1}{2\gamma}\left\|\frac{1}{L}\nabla_1 Q(\theta'|\theta') + \theta - \theta'\right\|^2}$ □

We place ourselves at iteration k of a penalized EM with the current value $\theta' = \theta_k \times \mathbb{R}^p$. //By the above lemma, we have for any $\gamma \in]0, 1/L]$ and $\theta \in \mathbb{R}^p$,

$$Q(\theta|\theta_k) - \text{pen}(\theta) \geqslant Q(\theta_k|\theta_k) + \frac{\gamma}{2}\left\|\nabla_1 Q(\theta_k|\theta_k)\right\|^2 - \frac{1}{2\gamma}\left\|\gamma\nabla_1 Q(\theta_k|\theta_k) + \theta - \theta_k\right\|^2 - \text{pen}(\theta)$$

Note that the equality case is achieve for $\theta = \theta_k$. So if we take the θ that achieve the maxima at the right-hands side of the inequality :

$$\begin{aligned} \theta_{k+1} &= \arg\max_{\theta \in \mathbb{R}^p}\left\{Q(\theta_k|\theta_k) + \frac{\gamma}{2}\left\|\nabla_1 Q(\theta_k|\theta_k)\right\|^2 - \frac{1}{2\gamma}\left\|\gamma\nabla_1 Q(\theta_k|\theta_k) + \theta - \theta_k\right\|^2 - \text{pen}(\theta)\right\} \\ &= \arg\min_{\theta \in \mathbb{R}^p}\left\{\frac{1}{2\gamma}\left\|\gamma\nabla_1 Q(\theta_k|\theta_k) + \theta - \theta_k\right\|^2 - \text{pen}(\theta)\right\} \\ &= prox_{\gamma, \text{pen}}\{\gamma\nabla_1 Q(\theta_k|\theta_k) + \theta_k\} \end{aligned} \tag{1}$$

So for any θ_{k+1} maximizing the right-hand side of the inequality, it holds the inequality, for

any $\theta \in \mathbb{R}^p$

$$\begin{aligned} Q(\theta_{k+1}|\theta_k) - \text{pen}(\theta_{k+1}) \geqslant & Q(\theta_k|\theta_k) + \frac{\gamma}{2}\left\|\nabla_1 Q(\theta_k|\theta_k)\right\|^2 \\ & - \frac{1}{2\gamma}\left\|\gamma\nabla_1 Q(\theta_k|\theta_k) + \theta_{k+1} - \theta_k\right\|^2 - \text{pen}(\theta_{k+1}) \\ \geqslant & Q(\theta_k|\theta_k) + \frac{\gamma}{2}\left\|\nabla_1 Q(\theta_k|\theta_k)\right\|^2 \\ & - \frac{1}{2\gamma}\left\|\gamma\nabla_1 Q(\theta_k|\theta_k) + \theta - \theta_k\right\|^2 - \text{pen}(\theta) \end{aligned}$$

In particular if we take $\theta = \theta_k$ for the far right-hand side of the inequality, we can conclude :

$$Q(\theta_{k+1}|\theta_k) - \text{pen}(\theta_{k+1}) \geqslant Q(\theta_k|\theta_k) - \text{pen}(\theta_k)$$

□

Remark *on lipshitz hypo*

6- Model inspirer de He et al. 2015

Rappel du modèle

Joint Model *final modeling*

For any genotype $1 \leqslant g \leqslant G$ and observation $1 \leqslant j \leqslant J$

$$\begin{cases} h_g(T_g|\mathcal{M}(T_g;\varphi_g), U_g) = h_0(T_g)\exp(\beta^T U_g + \alpha m(T_g;\varphi_g)) \\ Y_{g,j} = m(t_j;\varphi_g) + \epsilon_{g,j} \\ \varphi_g \sim \mathcal{N}(\mu,\Omega) \quad ; \quad \epsilon_{g,j} \sim \mathcal{N}(0,\sigma^2) \end{cases} \tag{2}$$

Pourquoi prendre $m(T_g;\varphi_g)$ alors que c'est une data observer (il me semble) ?
Si on a n_g observation pour le génotype g on pourrait remplacer cela par Y_{g,n_g} ?

Modèle par curiosité, inspirer de He et al. 2015 :
For the longitudinal outcome, we consider the following non-linear mixed effects model :

$$\begin{cases} Y_{g,j} = & m(t_j;\varphi_g) + \epsilon_{g,j} \\ \varphi_g = & \beta_1^T X_{1g} + b_g^T \Gamma_1 Z_{1g} \end{cases} \quad \epsilon_{g,j} \sim \mathcal{N}(0,\sigma^2)\ ;\ b_g \sim \mathcal{N}(0, I_q)$$

where $\beta_1 \in \mathbb{R}^p$ the first regression parameter, Γ_1 is a $q \times q$ lower triangular matrix.
For the survival outcome, we consider a frailty model, defined as follows :

$$h_g(T_g) = h_0(T_g)\exp(\beta_2^T x_{2g} + b_g^T \Gamma_2 z_{2g})$$

where h_0 is the baseline hazard function, $\beta_2 \in \mathbb{R}^p$ the second regression parameter, Γ_2 is a $q \times q$ lower triangular matrix.
The combine observation will be $\mathcal{O}_g = (Y_g, X_{1g}, Z_{1g}, T_g, X_{2g}, Y_{2g})$

Je sais pas si ce que je viens décrire est possible ni si cela à un sens en modélisation, en tout computationnellement parlant il y a un avantage indégniable : il n'y aura pas d'intégrale à calculer dans la likelihood !

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977). Publisher: [Royal Statistical Society, Wiley], pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875 (visited on 08/16/2022).

[2] Gersende Fort, Edouard Ollier, and Adeline Samson. *Stochastic Proximal Gradient Algorithms for Penalized Mixed Models*. 2017. DOI: 10.48550/ARXIV.1704.08891. URL: https://arxiv.org/abs/1704.08891.

[3] Zangdong He et al. "Simultaneous variable selection for joint models of longitudinal and survival outcomes: Variable Selection in Joint Models". en. In: *Biometrics* 71.1 (Mar. 2015), pp. 178–187. ISSN: 0006341X. DOI: 10.1111/biom.12221. URL: https://onlinelibrary.wiley.com/doi/10.1111/biom.12221 (visited on 10/25/2022).

Annexes

A) SAEM with gradient descent step

A-a) Standard SAEM in exponential context

Assuming we got observation written Y with unobserved latent variable Z in a model characterize by the parameter θ. We denote by $\mathcal{L}_{comp}$ the complete likelihood of Y. We place ourselves in the framework of exponential families and write the complete likelihood as follows: $\boxed{\mathcal{L}_{comp,Y}(\theta) = \langle \Phi(\theta); S(Y,Z) \rangle - \psi(\theta)}$. Then, we can write the A-step as :

$$S_{k+1} = (1-u_k)S_k + u_k S(Y, Z_{k+1})$$

The SAEM in the context of the exponential family can be written as follows:

Algorithm 4: SAEM for exponential family

Require : Number of iterations $K \geqslant 1$; $(u_k)_{k\geqslant 1}$ a sequence of step-size

Initialize Starting point $\theta_0 \in \mathbb{R}^d$; $S_0 = 0$; $H_0 = 0$

for $k = 1$ **to** K **do**

- **Step S :**
  simulate $Z^{(k)}$ according to the conditional distribution $f_Z(Z|Y,\theta_k)$
- **Step A :**
  update S_{k+1} according to a stochastic approximation :

$$S_{k+1} = (1-u_k)S_k + u_k S(Y, Z_{k+1})$$

- **Step M :**
  maximization : $\theta_{k+1} = \underset{\theta\in\Theta}{\arg\max} \{\langle \Phi(\theta); S_{k+1} \rangle - \psi(\theta)\}$

end

return $\hat{\theta} = \theta_K$

Where $(u_k)_{k\geqslant 1}$ is a step-size sequence such that $\forall k \in \mathbb{N}, u_k \in [0,1]$, $\sum_{k=1}^{\infty} u_k = \infty$ and $\sum_{k=1}^{\infty} u_k^2 < \infty$.

A-b) Adding gradient descent in maximization step

We assume that we can have, on the one hand, real parameters and, on the other hand, high dimensional parameters in θ. These last ones require a particular attention during the maximization that we wish to obtain by gradient descent. We denote by ν the parameters whose maximization is explicit and by β for the others : $\theta = (\nu, \beta) \in V \times B = \Theta \subset \mathbb{R}^{d+p}$.

To handle the high dimension we want to optimize the penalized likelihood as follows:

$$\hat{\theta} = (\hat{\nu}, \hat{\beta}) = \underset{\nu,\beta\in V\times B}{\arg\max}\, \mathcal{L}_{marg,Y}(\nu,\beta) - pen(\beta)$$

Where pen might be the LASSO penalty $pen(\beta) = \lambda \|\beta\|_1$ with λ a regularization parameter to be determined. We propose to compute the maximization step of the SAEM by performing a **two-step maximization**. We first calculate the explicit maxima over ν with β fixed equal to β_k :

$$\nu_{k+1} = \underset{\nu\in V}{\arg\max} \{\langle \Phi(\nu, \beta_k); S_{k+1} \rangle - \psi(\nu, \beta_k)\}$$

Then, using a specific method, we compute the maximization over the high-dimensional parameter β with ν being fixed equal to the updated value ν_{k+1} as :

$$\beta_{k+1} = \arg\max_{\beta \in B} \Big\{ \underbrace{\langle \Phi(\nu_{k+1}, \beta); S_{k+1} \rangle - \psi(\nu_{k+1}, \beta)}_{= Q_{k+1}(\beta)} - pen(\beta) \Big\}$$

This last step is achieve with a proximal gradient descent as follow :

Algorithm 5: Proximal Normalized Gradient Descent

Require : Number of iterations $K_{grad} \geqslant 1$, a function to optimize Q

Initialize $\beta_0 \in \mathbb{R}^p$ a starting point

for $k = 1$ **to** K_{grad} **do**

$\omega_k \leftarrow \beta_{k-1} - \gamma_k \dfrac{\nabla Q(\beta_{k-1})}{\|\nabla Q(\beta_{k-1})\|_2}$

$\beta_k \leftarrow prox_{\gamma_k pen}(\omega_k)$

end

return $\beta_{K_{grad}}$

The SAEM become the following Stochastic Approximation Expectation Gradient Descent Maximization

Algorithm 6: SAEGDM

Require : Number of iterations $K \geqslant 1$; $(u_k)_{k \geqslant 1}$ a sequence of step-size

Initialize Starting point $\theta_0 \in \mathbb{R}^d$; $S_0 = 0$; $H_0 = 0$

for $k = 1$ **to** K **do**

• **Step S :**

simulate $Z^{(k)}$ according to the conditional distribution $f_Z(Z|Y, \theta_k)$

• **Step A :**

update S_{k+1} and H_{k+1} according to a stochastic approximation :

$$S_{k+1} = (1 - u_k) S_k + u_k S(Y, Z_{k+1})$$

• **Step M :**

• maximization : $\nu_{k+1} = \arg\max_{\nu \in V} \{ \langle \Phi(\nu, \beta_k); S_{k+1} \rangle - \psi(\nu, \beta_k) \}$

• gradient descent on Q_{k+1}:

$$\beta_{k+1} = \arg\max_{\beta \in B} \Big\{ \underbrace{\langle \Phi(\nu_{k+1}, \beta); S_{k+1} \rangle - \psi(\nu_{k+1}, \beta)}_{= Q_{k+1}(\beta)} - pen(\beta) \Big\}$$

end

return $\hat{\theta} = (\hat{\nu}, \hat{\beta}) = (\nu_K, \beta_K)$