1 Le modèle

On a un modèle génératif M complexe qui se décompose en trois morceaux $M = (M_1, M_2, M_3)$ comme suit :

$$M_1 : \quad Y|X^{I_1}, X^{I_2}, Z \sim X^{I_1}A_{I_1} + X^{I_2}A_{I_2} + \varepsilon \tag{1}$$
$$M_2 : \quad X^{I_2}|X^{I_1}, Z \sim X^{I_1}B_{I_1}^{I_2} + \varepsilon_{X^{I_2}} \tag{2}$$
$$M_3 : \quad X^{I_1}|Z \sim GM(\alpha) \tag{3}$$

On pose $\theta = (\alpha, \Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2})$ Pour un Z donné on a donc

$$\begin{aligned} P_{M,Z}(Y, X^{I_1}, X^{I_2}|\theta) &= P_{M,Z}(X^{I_1}|\theta)P_{M,Z}(X^{I_2}|X^{I_1};\theta)P_{M,Z}(Y|X^{I_2}, X^{I_1};\theta) &(4)\\ &= P_{M_3,Z}(X^{I_1}|\alpha)P_{M_1,M_2,Z}(X^{I_2}|X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2})P_{M_1,M_2,Z}(Y|X^{I_2}, X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2}) &(5)\end{aligned}$$

$P_{M_3,Z}(X^{I_1}|\alpha)$ vit tout seul et est estimé par Mixmod. On sait également que

$$\begin{aligned} P_{M_2,Z}(X^{I_2}|X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2}) &= P_{M_2,Z}(X^{I_2}|X^{I_1};\Sigma_{X^{I_2}}, B_{I_1}^{I_2}) \text{ et} &(6)\\ P_{M_1,Z}(Y|X^{I_2}, X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2}) &= P_{M_1,Z}(Y|X^{I_2}, X^{I_1};\Sigma, A_{I_1}, A_{I_2}) &(7)\end{aligned}$$

De manière générale les gens s'arrêtent là, et obtiennent le modèle complet... qui ne tient pas compte de la structure. Mais nous, nous ne sommes pas juste sous M_2 pour les sous-régression ni juste sous M_1 pour Y. Donc la dépendance persiste (c'est ce qui permet que la contrainte aie une utilité, sinon elle serait inutile).

Du coup, il faut définir proprement $P_{M_1,M_2,Z}(X^{I_2}|X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2})$ et $P_{M_1,M_2,Z}(Y|X^{I_2}, X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2})$. Or il semble que ces vraisemblances ne soient pas explicites.

Vu que $P_{M_1,M_2,Z}(X^{I_2}|X^{I_1};\theta) \neq P_{M_2,Z}(X^{I_2}|X^{I_1};\theta)$ et $P_{M_1,M_2,Z}(Y|X^{I_2}, X^{I_1};\theta) \neq P_{M_1,Z}(Y|X^{I_2}, X^{I_1};\theta)$, on se limite donc à une décomposition partielle de la vraisemblance et on obtient :

$$P_{M,Z}(Y, X^{I_1}, X^{I_2}|\theta) = P_{M_3,Z}(X^{I_1}|\alpha)P_{M_1,M_2,Z}(Y, X^{I_2}|X^{I_1};\Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}) \tag{8}$$

Pour alléger les notations, on va pour la suite de cette partie considérer les individus séparément (sans perte de généralité vu qu'ils sont $i.i.d$). On a alors Y scalaire et X vecteur de taille p. Le second membre suit une loi normale mulivariée :

$$(Y, X^{I_2}|X^{I_1}) \sim \mathcal{N}\left(\begin{pmatrix} X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}; \begin{pmatrix} \sigma_Y^2 + \sum_{i\in I_2}(\sigma_i^2 a_i^2) & \dots & a_j\sigma_j^2 & \dots \\ \vdots & \ddots & 0 & 0 \\ a_j\sigma_j^2 & 0 & \sigma_j^2 & 0 \\ \vdots & 0 & 0 & \ddots \end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix} X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}; \begin{pmatrix} \sigma_Y^2 + \sum_{i\in I_2}(\sigma_i^2 a_i^2) & \dots a_j\sigma_j^2 \dots \\ \vdots & \\ a_j\sigma_j^2 & \Sigma_X^{I_2} \\ \vdots & \end{pmatrix}\right) \tag{9}$$

On note $\bar{\Sigma}$ la matrice de variance-covariance correspondante Avec par conséquent une vraisemblance de la forme :

$$P_{M_1,M_2,Z}(Y, X^{I_2}|X^{I_1}; \Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}) = \mathcal{L}_{M_1,M_2,Z}(\Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}); Y, X^{I_2}|X^{I_1}) \tag{10}$$

$$= \frac{1}{(2\pi)^{\frac{p_2+1}{2}}|\bar{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\begin{pmatrix} Y - X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}^t \bar{\Sigma}^{-1} \begin{pmatrix} Y - X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}\right) \tag{11}$$

avec $|\bar{\Sigma}|$ le déterminant de $\bar{\Sigma}$.

On passe à la log-vraisemblance :

$$L_{M_1,M_2,Z}(\Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}; Y, X^{I_2}|X^{I_1}) = -\frac{1}{2}\left[(p_2+1)\ln(2\pi) + \ln(|\bar{\Sigma}|) + \begin{pmatrix} Y - X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}^t \bar{\Sigma}^{-1} \begin{pmatrix} Y - X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}\right] \tag{12}$$

2 estimation

On veut maximiser la vraisemblance sous la contrainte $A_{I_1} + B_{I_1}^{I_2}A_{I_2} = \hat{\hat{A}}_{I_1}$ où $\hat{\hat{A}}_{I_1}$ a été préalablement calculé lors de l'étape explicative. Cette contrainte n'est pas linéaire mais devient linéaire si on se limite à A ou à B. On va donc procéder par optimisation alternée, en maximisant la vraisemblance pour un B fixé puis en recommençant avec le A obtenu, et ainsi de suite.La contrainte étant d'égalité, on utilisera le théorème de Lagrange. La partie de la vraisemblance relative à X^{I_1} ne dépend pas de la contrainte ni de l'autre membre et on la maximise donc en amont par Mixmod.

On commence par expliciter la matrice $\bar{\Sigma}^{-1}$

$$\bar{\Sigma} = \begin{pmatrix} E & C^t \\ C & \Sigma_{X^{I_2}} \end{pmatrix} \tag{13}$$

$$E = \sigma_Y^2 + \sum_{i \in I_2}(\sigma_i^2 A_i^2) = \sigma_Y^2 + A_{I_2}^t \Sigma_{X^{I_2}} A_{I_2} \text{ scalaire} \tag{14}$$

$$C = \begin{pmatrix} \vdots \\ a_i\sigma_i^2 \\ \vdots \end{pmatrix} = \Sigma_{X^{I_2}} A_{I_2} \text{ de taille } p_2 \times 1 \tag{15}$$

$$C^t = A_{I_2}^t \Sigma_{X^{I_2}}^t = A_{I_2}^t \Sigma_{X^{I_2}} \text{ de taille } 1 \times p_2 \tag{16}$$

On a donc à l'aide du complément de Schur :

$$\bar{\Sigma}^{-1} = \begin{pmatrix} [E - C^t\Sigma_{X^{I_2}}^{-1}C]^{-1} & \cdot \\ -\Sigma_{X^{I_2}}^{-1}C(E - C^t\Sigma_{X^{I_2}}^{-1}C)^{-1} & [\Sigma_{X^{I_2}}^{-1} + \Sigma_{X^{I_2}}^{-1}C(E - C^t\Sigma_{X^{I_2}}^{-1}C)^{-1}C^t\Sigma_{X^{I_2}}^{-1}] \end{pmatrix} \text{ matrice symétrique} \tag{17}$$

Le complément de Schur est récurrent, on le note $S = E - C^t\Sigma_{X^{I_2}}^{-1}C$. Après calculs on obtient $S^{-1} = \frac{1}{\sigma_Y^2}$ On obtient alors :

$$-\Sigma_{X^{I_2}}^{-1}C(E - C^t\Sigma_{X^{I_2}}^{-1}C)^{-1} = -\Sigma_{X^{I_2}}^{-1}CS^{-1} = -\Sigma_{X^{I_2}}^{-1}\Sigma_{X^{I_2}}A_{I_2}S^{-1} = -\frac{1}{\sigma_Y^2}A_{I_2} \text{ de taille } p_2 \times 1 \tag{18}$$

et enfin :

$$[\Sigma_{X^{I_2}}^{-1} + \Sigma_{X^{I_2}}^{-1}C(E - C^t\Sigma_{X^{I_2}}^{-1}C)^{-1}C^t\Sigma_{X^{I_2}}^{-1}] = \Sigma_{X^{I_2}}^{-1} + \frac{1}{\sigma_Y^2}A_{I_2}A_{I_2}^t\Sigma_{X^{I_2}}\Sigma_{X^{I_2}}^{-1} = \Sigma_{X^{I_2}}^{-1} + \frac{1}{\sigma_Y^2}A_{I_2}A_{I_2}^t \text{ matrice symétrique pleine} \tag{19}$$

On a donc :

$$\bar{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\sigma_Y^2} & -\frac{1}{\sigma_Y^2}A_{I_2}^t \\ -\frac{1}{\sigma_Y^2}A_{I_2} & [\Sigma_{X^{I_2}}^{-1} + \frac{1}{\sigma_Y^2}A_{I_2}A_{I_2}^t] \end{pmatrix} \text{ matrice symétrique pleine} \tag{20}$$

On note $\Sigma_{X^{I_2}(-k)}$ la matrice $\Sigma_X^{I_2}$ dépourvue de la ligne$k$ et de la colonne k. On note $\bar{\Sigma}(-1,-k)$ la matrice $\bar{\Sigma}$ dépourvue de sa prémière colonne et de sa ligne k. On a alors

$$\begin{aligned}
|\bar{\Sigma}| &= (\sigma_Y^2 + A_{I_2}^t\Sigma_{X^{I_2}}A_{I_2})|\Sigma_X^{I_2}| + \sum_{k=1}^{p_2}(-1)^k A_{I_2(k)}\sigma_{I_2(k)}^2|\bar{\Sigma}(-1,-(k+1))| \text{ développement 1 colonne} && (21)\\
&= (\sigma_Y^2 + A_{I_2}^t\Sigma_{X^{I_2}}A_{I_2})|\Sigma_X^{I_2}| + \sum_{k=1}^{p_2}(-1)^k A_{I_2(k)}\sigma_{I_2(k)}^2[(-1)^{k+1}A_{I_2(k)}\sigma_{I_2(k)}^2\frac{1}{\sigma_{I_2(k)}^2}\prod_{j\in I_2}\sigma_j^2] \text{ developpements k colonne} && (22)\\
&= \sigma_Y^2\prod_{j\in I_2}\sigma_j^2 + \sum_{k=1}^{p_2}A_{I_2(k)}^2\sigma_{I_2(k)}^2\prod_{j\in I_2}\sigma_j^2 - \sum_{k=1}^{p_2}A_{I_2(k)}^2\sigma_{I_2(k)}^2\prod_{j\in I_2}\sigma_j^2 && (23)\\
&= \sigma_Y^2\prod_{j\in I_2}\sigma_j^2 && (24)
\end{aligned}$$

On peut donc calculer plus finement la log-vraisemblance :

$$\begin{aligned}
L_{M,Z}(\Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}); Y, X^{I_2}|X^{I_1}) &= -\frac{1}{2}\left[(p_2+1)\ln(2\pi) + \ln(|\bar{\Sigma}|) + \begin{pmatrix} Y - X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}^t \bar{\Sigma}^{-1}\begin{pmatrix} Y - X^{I_1}(A_{I_1} + B_{I_1}^{I_2}A_{I_2}) \\ (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t \end{pmatrix}\right] && (25)\\
&= \dots && \\
&= -\frac{1}{2}\left[(p_2+1)\ln(2\pi) + \ln(|\bar{\Sigma}|) + (X^{I_2} - X^{I_1}B_{I_1}^{I_2})\Sigma_{X^{I_2}}^{-1}(X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t + \frac{1}{\sigma_Y^2}(Y - XA)^2\right] && (26)
\end{aligned}$$

2.1 initialisation

On commence par choisir une valeur de B. On pourrait en prendre une arbitraire, mais on choisit de commencer par le B obtenu par estimation indépendante des sous-régressions (par maximisation de $P_{M_2,Z}(X^{I_2}|X^{I_1}; \Sigma_{X^{I_2}}, B_{I_1}^{I_2})$). Cela peut se faire par moindres carrés classiques appliqués de manière indépendante à chacune des sous-régressions.

2.2 étape A

pour cette étape, les paramètres $\Sigma_{X^{I_2}}$ et $B_{I_1}^{I_2}$ sont fixés. On se ramène à la maximisation de

$$\begin{aligned} \max_A f(A) &= \max_A L_{M_1,M_2,Z}(\Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}; Y, X^{I_2}|X^{I_1}) \quad (27) \\ s.c.\ \ A_{I_1} + B_{I_1}^{I_2} A_{I_2} - \hat{\hat{A}}_{I_1} &= 0 \quad (28) \end{aligned}$$

En notant $\psi(A) = A_{I_1} + B_{I_1}^{I_2} A_{I_2} - \hat{\hat{A}}_{I_1} = B_{I_1} A - \hat{\hat{A}}_{I_1}$ la contrainte, le théorème de Lagrange permet de dire qu'il existe un unique λ pour lequel résoudre notre maximisation sous contrainte revient à résoudre :

$$\bigtriangledown f(A, \sigma_Y^2) + \lambda \bigtriangledown \psi(A) = 0 \quad (29)$$

On note $g(A, \lambda) = f(A, \sigma_Y^2) + \lambda\psi(A)$.

pour $i \in I$:

$$\begin{aligned} \frac{\partial g(A, \lambda, \sigma_Y^2)}{\partial A_i} &= \frac{\partial}{\partial A_i}\left(-\frac{1}{2\sigma_Y^2}(Y - XA)^t(Y - XA) + \lambda(B_{I_1}A - \hat{\hat{A}}_{I_1})\right) \\ &= -\frac{1}{2\sigma_Y^2}\left(-2(X^i)^t Y + 2(X^i)^t XA\right) + \lambda B_{I_1}^i \\ &= \frac{1}{\sigma_Y^2}(X^i)^t(Y - XA) + \lambda B_{I_1}^i \quad (30) \end{aligned}$$

On a pour $i \in I_1$: $\lambda B_{I_1}^i = \lambda_i$ et pour $i \in I_2$: $\lambda B_{I_1}^i = \sum_{j \in I_1} \lambda_j B_{j,i}$

et $\forall i \in I_1$ on a :

$$\frac{\partial g(A, \lambda, \sigma_Y^2)}{\partial \lambda_i} = \frac{\partial \lambda\psi(A)}{\partial \lambda_i} = (B_i A - \hat{\hat{A}}_i) = A_i + \sum_{j \in I_2} B_{i,j} A_j - \hat{\hat{A}}_i \quad (31)$$

$$\begin{aligned} \frac{\partial g(A, \lambda, \sigma_Y^2)}{\partial \sigma_Y} &= -\frac{1}{2}\frac{\partial}{\partial \sigma_Y}\left(\ln(\sigma_Y^2 \prod_{j \in I_2} \sigma_j^2) + \frac{1}{\sigma_Y^2}(Y - XA)^t(Y - XA)\right) = -\frac{1}{2}\left[\frac{2}{\sigma_Y} - \frac{2}{\sigma_Y^3}(Y - XA)^t(Y - XA)\right] \\ &= \frac{1}{\sigma_Y^3}(Y - XA)^t(Y - XA) - \frac{1}{\sigma_Y} \quad (32) \end{aligned}$$

On revient maintenant à la notation matricielle Y vecteur de taille n et X matrice de taille $n \times p$. On a donc un système à résoudre de la forme

(individus iid+log-vraisemblance)

$$A_{I_1} + B_{I_1}^{I_2} A_{I_2} - \hat{\hat{A}}_{I_1} = 0 \quad (33)$$

$$\frac{1}{\sigma_Y^2}(X^{I_1})^t(Y - X^{I_1}A_{I_1} - X^{I_2}A_{I_2}) + \lambda^t = 0 \quad (34)$$

$$\frac{1}{\sigma_Y^2}(X^{I_2})^t(Y - X^{I_1}A_{I_1} - X^{I_2}A_{I_2}) + (B_{I_1}^{I_2})^t\lambda^t = 0 \quad (35)$$

$$\frac{1}{\sigma_Y^3}(Y - XA)^t(Y - XA) - \frac{n}{\sigma_Y} = 0 \quad (36)$$

On annule la différentielle de la vraisemblance, qui est la somme des vraisemblances des individus donc par linéarité de la différentielle on annule la somme des différentielles des vraisemblances des individus.

ce système s'écrit également :

$$A_{I_1} = \hat{\hat{A}}_{I_1} - B_{I_1}^{I_2} A_{I_2} \quad (37)$$

$$\frac{1}{\sigma_Y^2}(X^{I_1})^t(Y - X^{I_1}(\hat{\hat{A}}_{I_1} - B_{I_1}^{I_2}A_{I_2}) - X^{I_2}A_{I_2}) + \lambda^t = 0 \quad (38)$$

$$\frac{1}{\sigma_Y^2}(X^{I_2})^t(Y - X^{I_1}(\hat{\hat{A}}_{I_1} - B_{I_1}^{I_2}A_{I_2}) - X^{I_2}A_{I_2}) + (B_{I_1}^{I_2})^t\lambda^t = 0 \quad (39)$$

$$\sigma_Y^2 = \frac{1}{n}(Y - XA)^t(Y - XA) \quad (40)$$

On en déduit :

$$\lambda^t = -\frac{1}{\sigma_Y^2}(X^{I_1})^t(Y - X^{I_1}A_{I_1} - X^{I_2}A_{I_2}) \quad (41)$$

$$\text{d'où} \quad 0 = \sigma_Y^2\frac{1}{\sigma_Y^2}(X^{I_2})^t(Y - X^{I_1}(\hat{\hat{A}}_{I_1} - B_{I_1}^{I_2}A_{I_2}) - X^{I_2}A_{I_2}) - \sigma_Y^2\frac{1}{\sigma_Y^2}(B_{I_1}^{I_2})^t(X^{I_1})^t(Y - X^{I_1}A_{I_1} - X^{I_2}A_{I_2}) \quad (42)$$

$$0 = (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t(Y - X^{I_1}\hat{\hat{A}}_{I_1} - (X^{I_2} - X^{I_1}B_{I_1}^{I_2})A_{I_2}) \quad (43)$$

$$\text{d'où} \quad (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t(Y - X^{I_1}\hat{\hat{A}}_{I_1}) = (X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t(X^{I_2} - X^{I_1}B_{I_1}^{I_2})A_{I_2} \quad (44)$$

$$\text{et enfin} \quad \hat{A}_{I_2} = \left((X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t(X^{I_2} - X^{I_1}B_{I_1}^{I_2})\right)^{-1}(X^{I_2} - X^{I_1}B_{I_1}^{I_2})^t(Y - X^{I_1}\hat{\hat{A}}_{I_1}) \quad (45)$$

$$\text{puis } \hat{A}_{I_1} = \hat{\hat{A}}_{I_1} - B_{I_1}^{I_2}A_{I_2} \quad (46)$$

On remarque que $(X^{I_2} - X^{I_1}B_{I_1}^{I_2}) = \varepsilon_{X^{I_2}}$ résidu des sous-régressions et que $Y - X^{I_1}\hat{\hat{A}}_{I_1} = \tilde{\varepsilon}$ résidu du modèle explicatif. On se contente donc en fait

d'estimer les résidus du modèle explicatif par une seconde régression linéaire.

$$Y - X^{I_1}\hat{\hat{A}}_{I_1} = \varepsilon_{X^{I_2}} A_{I_2} + \varepsilon_Y \tag{47}$$

σ_Y se calcule tout simplement comme l'écart-type empirique du résidu de la régression : $\hat{\sigma}_Y^2 = \frac{1}{n}(Y - X\hat{A})^t(Y - X\hat{A})$.

De ce fait, on peut utiliser tout estimateur et méthode de sélection propre à la régression linéaire en l'appliquant à ces données modifiées. Ce modèle est appelé modèle prédictif. L'idée est maintenant d'optimiser B sachant A pour ensuite venir recalculer ce modèle, et ainsi de suite jusqu'à convergence vers le maximum de vraisemblance de la loi jointe du modèle (M).

2.3 étape B

Pour cette étape, les paramètres Σ, A_{I_1} et A_{I_2} sont fixés (ici aussi, la connaissance de A permet d'obtenir le Σ associé). Comme pour l'étape A, on va résoudre :

$$\bigtriangledown f(B) + \lambda \bigtriangledown \psi(B) = 0 \tag{48}$$

où $f(B) = L_{M_1,M_2,Z}(\Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}); Y, X^{I_2}|X^{I_1})$

$$g(B, \lambda, \Sigma_{X^{I_2}}) = f(B, \Sigma_{X^{I_2}}) + \lambda\psi(B) = 0 \tag{49}$$

On note $I_1^j = \{i \in I_1 | Z_{i,j} \neq 0\}$. Alors $\forall j \in I_2, \forall i \in I_1^j$:

$$\begin{aligned}
\frac{\partial g(B, \lambda, \Sigma_{X^{I_2}})}{\partial B_{i,j}} &= \frac{\partial}{\partial B_{i,j}} \sum_{k=1}^{n} \left(-\frac{1}{2}[(X_k^{I_2} - X_k^{I_1} B_{I_1}^{I_2}) \Sigma_{X^{I_2}}^{-1} (X_k^{I_2} - X_k^{I_1} B_{I_1}^{I_2})^t] \right) + \frac{\partial}{\partial B_{i,j}} \left(\lambda (B_{I_1} A - \hat{\hat{A}}) \right) && (50) \\
&= -\frac{1}{2} \frac{\partial}{\partial B_{i,j}} \left(\sum_{l \in I_2} \frac{1}{\sigma_l^2} (X^l - \sum_{k \in I_1} X^k B_{k,l})^t (X^l - \sum_{k \in I_1} X^k B_{k,l}) \right) + \lambda_i A_j && (51) \\
&= \frac{1}{\sigma_j^2} (X^j - \sum_{k \in I_1} X^k B_{k,j})^t X^i + \lambda_i A_j && (52) \\
&= \frac{1}{\sigma_j^2} (X^j - \sum_{k \in I_1^j} X^k B_{k,j})^t X^i + \lambda_i A_j && (53) \\
\frac{\partial g(B, \lambda, \Sigma_{X^{I_2}})}{\partial B_{I_1^j}^j} &= \frac{1}{\sigma_j^2} (X^{I_1^j})^t (X^j - X^{I_1^j} B_{I_1^j}^j) + A_j \lambda_{I_1^j}^t && (54)
\end{aligned}$$

avec $\lambda_{I_1^j}$ qui désigne le vecteur des λ_i associés aux élements de I_1^j. et pour tout $i \in I_1$:

$$\frac{\partial g(B,\lambda,\Sigma_{X^{I_2}})}{\partial \lambda_i} = \frac{\partial \lambda \psi(B)}{\partial \lambda_i} = (B_i A - \hat{\hat{A}}_i) = A_i + \sum_{j \in I_2} B_{i,j} A_j - \hat{\hat{A}}_i = 0 \tag{55}$$

$$\begin{aligned}
\frac{\partial g(B,\lambda,\Sigma_{X^{I_2}})}{\partial \sigma_j} &= -\frac{1}{2}\left[\frac{\partial}{\partial \sigma_j}\left(\sum_{i=1}^{n}\sum_{k\in I_2}\frac{1}{\sigma_k^2}(X_i^k - X_i^{I_1}B_{I_1}^k)^2\right) + n\frac{\partial}{\partial \sigma_j}\ln(\sigma_Y^2\prod_{j\in I_2}\sigma_j^2)\right] && (56)\\
&= -\frac{1}{2}\left[-\frac{2}{\sigma_j^3}\sum_{i=1}^{n}(X_i^j - X_i^{I_1}B_{I_1}^j)^2 + n\frac{2}{\sigma_j}\right] = n\frac{1}{\sigma_j} - \frac{1}{\sigma_j^3}(X^j - X^{I_1}B_{I_1}^j)^t(X^j - X^{I_1}B_{I_1}^j) && (57)\\
&= n\frac{1}{\sigma_j} - \frac{1}{\sigma_j^3}(X^j - X^{I_1^j}B_{I_1^j}^j)^t(X^j - X^{I_1^j}B_{I_1^j}^j) && (58)
\end{aligned}$$

Et donc on obtient le système :

$$\begin{aligned}
\forall j \in I_2 :\ \frac{1}{\sigma_j^2}(X^{I_1^j})^t(X^j - X^{I_1^j}B_{I_1^j}^j) + A_j\lambda_{I_1^j}^t &= 0 && (59)\\
A_{I_1} + \sum_{j\in I_2} A_j B_{I_1}^j - \hat{\hat{A}}_{I_1} &= 0 && (60)\\
\forall j \in I_2 : \sigma_j^2 &= \frac{1}{n}(X^j - X^{I_1^j}B_{I_1^j}^j)^t(X^j - X^{I_1^j}B_{I_1^j}^j) && (61)
\end{aligned}$$

3 Concavité de la vraisemblance

On s'assure de la concavité de la vraisemblance en B, σ pour confirmer l'importance de la résolution du système.
Comme on vaut garder la contrainte, on vérifie plutôt la concavité de g et donc on ajoute la paramètre λ dans les variables.
On décompose la hessienne de la manière suivante :

$$H = \begin{pmatrix} H_1 & H_2 & H_3 \\ H_4 & H_5 & H_6 \\ H_7 & H_8 & H_9 \end{pmatrix} = \begin{pmatrix} \frac{\partial^2}{\partial\sigma_i\partial\sigma_j} & \frac{\partial^2}{\partial B^i_{I_1^i}\partial\sigma_j} & \frac{\partial^2}{\partial\lambda_i\partial\sigma_j} \\ \frac{\partial^2}{\partial\sigma_i\partial B^j_{I_1^j}} & \frac{\partial^2}{\partial B^i_{I_1^i}\partial B^j_{I_1^j}} & \frac{\partial^2}{\partial\lambda_i\partial B^j_{I_1^j}} \\ \frac{\partial^2}{\partial\sigma_i\partial\lambda_j} & \frac{\partial^2}{\partial B^i_{I_1^i}\partial\lambda_j} & \frac{\partial^2}{\partial\lambda_i\partial\lambda_j} \end{pmatrix} \tag{62}$$

$$\frac{\partial g}{\partial \sigma_j} = \frac{n}{\sigma_j} - \frac{1}{\sigma_j^3}(X^j - X^{I_1^j}B_{I_1^j}^j)^t(X^j - X^{I_1^j}B_{I_1^j}^j) \tag{63}$$

donc, pour H_1 :

$$\frac{\partial^2 g}{\partial \sigma_i \partial \sigma_j} = 0 \text{ si } i \neq j \tag{64}$$

$$\frac{\partial^2 g}{\partial \sigma_j \partial \sigma_j} = -\frac{n}{\sigma_j^2} + \frac{1}{\sigma_j^4}(X^j - X^{I_1^j} B^j_{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) \tag{65}$$

H_1 est donc diagonale.

Pour H_2 :

$$\frac{\partial^2 g}{\partial B^i_{I_1^i} \partial \sigma_j} = 0 \text{ si } i \neq j \tag{66}$$

$$\frac{\partial^2 g}{\partial B^j_{I_1^j} \partial \sigma_j} = -\frac{2}{\sigma_j^3}\left((X^{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j})\right)^t \tag{67}$$

Donc H_2 est diagonale par blocs (vecteurs lignes). Et pour H_3 :

$$\frac{\partial^2 g}{\partial \lambda_i \partial \sigma_j} = 0 \tag{68}$$

On a aussi

$$\frac{\partial g}{\partial B^j_{I_1^j}} = \frac{1}{\sigma_j^2}(X^{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) + \lambda^t_{I_1^j} A_j \tag{69}$$

Donc pour H_4 :

$$\frac{\partial^2 g}{\partial \sigma_i \partial B^j_{I_1^j}} = 0 \text{ si } i \neq j \tag{70}$$

$$\frac{\partial^2 g}{\partial \sigma_j \partial B^j_{I_1^j}} = -\frac{2}{\sigma_j^3}(X^{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) \tag{71}$$

H_4 matrice diagonale par blocs (vecteurs colonne) et $H_4 = H_2^t$

Pour H_5 :

$$\frac{\partial^2 g}{\partial B^i_{I_1^i} \partial B^j_{I_1^j}} = 0 \text{ si } i \neq j \tag{72}$$

$$\frac{\partial^2 g}{\partial B^j_{I_1^j} \partial B^j_{I_1^j}} = -\frac{1}{\sigma_j^2}(X^{I_1^j})^t X^{I_1^j} \tag{73}$$

et pour H_6 :

$$\frac{\partial^2 g}{\partial \lambda_i \partial B^j_{I_1^j}}(k) = A_j \text{ si } i \in I_1^j \text{ et } k = i \quad (74)$$

$$= 0 \text{ sinon} \quad (75)$$

Donc H_6 est diagonal

On a enfin

$$\frac{\partial g}{\partial \lambda} = A_{I_1} + B^{I_2}_{I_1} A_{I_2} \quad (76)$$

Donc $H_7 = 0$, $H_9 = 0$ et $H_8 = (H_6)^t$

4 Methode de Newton

On veut résoudre le système non linéaire suivant :

$$f1 : \forall j \in I_2 : \frac{1}{\sigma_j^2}(X^{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) + A_j \lambda^t_{I_1^j} = 0 \quad (77)$$

$$f2 : A_{I_1} + \sum_{j \in I_2} A_j B^j_{I_1} - \hat{\hat{A}}_{I_1} = 0 \quad (78)$$

$$f3 : \forall j \in I_2 : \sigma_j^2 - \frac{1}{n}(X^j - X^{I_1^j} B^j_{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) = 0 \quad (79)$$

On va utiliser l'algorithme de Newton On commence donc par calculer la jacobienne associée :

$$\mathcal{J} = \begin{pmatrix} \mathcal{J}_1 & \mathcal{J}_2 & \mathcal{J}_3 \\ \mathcal{J}_4 & \mathcal{J}_5 & \mathcal{J}_6 \\ \mathcal{J}_7 & \mathcal{J}_8 & \mathcal{J}_9 \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial B_i^j} & \frac{\partial f_1}{\partial \lambda_i} & \frac{\partial f_1}{\partial \sigma_j} \\ \frac{\partial f_2}{\partial B_i^j} & \frac{\partial f_2}{\partial \lambda_i} & \frac{\partial f_2}{\partial \sigma_j} \\ \frac{\partial f_3}{\partial B_i^j} & \frac{\partial f_3}{\partial \lambda_i} & \frac{\partial f_3}{\partial \sigma_j} \end{pmatrix} \quad (80)$$

On note $p_Z = \sum i \in I_1, j \in I_2 Z_{i,j}$ le nombre de paramètres à estimer dans B et $I_Z = \{1, \ldots, p_Z\}$. Les dimensions des blocs sont les suivantes :

- $\mathcal{J}_1 : p_Z \times p_Z$
- $\mathcal{J}_2 : p_Z \times p_1$
- $\mathcal{J}_3 : p_Z \times p_2$
- $\mathcal{J}_4 : p_1 \times p_Z$
- $\mathcal{J}_5 : p_1 \times p_1$

– $\mathcal{J}_6 : p_1 \times p_2$
– $\mathcal{J}_7 : p_2 \times p_Z$
– $\mathcal{J}_8 : p_2 \times p_1$
– $\mathcal{J}_9 : p_2 \times p_2$

On remarque tout de suite que $\mathcal{J}_5, \mathcal{J}_6, \mathcal{J}_8$ sont nuls

$$\mathcal{J} = \begin{pmatrix} \frac{\partial f_1}{\partial B_i^j} & \frac{\partial f_1}{\partial \lambda_i} & \frac{\partial f_1}{\partial \sigma_j} \\ \frac{\partial f_2}{\partial B_i^j} & 0 & 0 \\ \frac{\partial f_3}{\partial B_i^j} & 0 & \frac{\partial f_3}{\partial \sigma_j} \end{pmatrix} \tag{81}$$

On note $\bar{Z}$ la matrice $p_Z \times 2$ des indices non nuls de Z ordonnés par colonne. On a donc en colonne $k \in I_Z$ les dérivées partielles

$$\frac{\partial}{\partial B_{\bar{Z}_{k,1}}^{\bar{Z}_{k,2}}}$$

pour $\mathcal{J}_9$

$$\begin{aligned} \forall (j,k) \in I_2 \times I_2 : \ \mathcal{J}_9(j,k) &= \frac{\partial}{\partial \sigma_k} \left(\sigma_j^2 - \frac{1}{n} (X^j - X^{I_1^j} B^j_{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) \right) && (82) \\ &= 0 \text{ si } k \neq j && (83) \\ &= 2\sigma_k \text{ sinon} && (84) \end{aligned}$$

matrice diagonale.

pour $\mathcal{J}_7$

$$\forall (j,k) \in I_2 \times I_Z, \ \mathcal{J}_7(j,k) = \frac{\partial}{\partial B_{\bar{Z}_{k,1}}^{\bar{Z}_{k,2}}} \left(\sigma_j^2 - \frac{1}{n} (X^j - X^{I_1^j} B^j_{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) \right) = 0 \text{ si } \bar{Z}_{k,2} \neq j \tag{85}$$

donc $\mathcal{J}_7$ est diagonale par blocs. Et pour $\bar{Z}_{k,2} = j$:

$$\begin{aligned}
\mathcal{J}_7(j,k) &= \frac{\partial}{\partial B^j_{\bar{Z}_{k,1}}} \left(\sigma_j^2 - \frac{1}{n}(X^j - X^{I_1^j} B^j_{I_1^j})^t (X^j - X^{I_1^j} B^j_{I_1^j}) \right) && (86)\\
&= -\frac{1}{n} \frac{\partial}{\partial B^j_{\bar{Z}_{k,1}}} \left((X^j)^t X^j - 2(X^j)^t X^{I_1^j} B^j_{I_1^j} + (X^{I_1^j} B^j_{I_1^j})^t X^{I_1^j} B^j_{I_1^j}) \right) && (87)\\
&= -\frac{1}{n} \frac{\partial}{\partial B^j_{\bar{Z}_{k,1}}} \left(-2(X^j)^t \sum_{i \in I_1^j} X^i B_i^j + \sum_{l=1}^{n} \sum_{i \in I_1^j} (X_l^i B_i^j)^2 \right) && (88)\\
&= -\frac{1}{n}(-2(X^j)^t X^{\bar{Z}_{k,1}} + 2B^j_{\bar{Z}_{k,1}} (X^{\bar{Z}_{k,1}})^t X^{\bar{Z}_{k,1}}) && (89)\\
&= \frac{2}{n}(X^j - X^{\bar{Z}_{k,1}} B^j_{\bar{Z}_{k,1}})^t X^{\bar{Z}_{k,1}} && (90)
\end{aligned}$$

On passe à $\mathcal{J}_4$:

$$\begin{aligned}
\forall (i,k) \in I_1 \times I_Z,\ \mathcal{J}_4(\tilde{i},k) &= \frac{\partial}{\partial B^{\bar{Z}_{k,2}}_{\bar{Z}_{k,1}}} \left(A_i + \sum_{j \in I_2} A_j B_i^j - \hat{\hat{A}}_i \right) = 0 \text{ si } \bar{Z}_{k,1} \neq i && (91)\\
&= A_{\bar{Z}_{k,2}} \text{ si } \bar{Z}_{k,1} = i && (92)
\end{aligned}$$

où $\tilde{i}$ est la position de i dans I_1. Matrice par blocs diagonaux (blocs alignés côte à côte, un par sous-régression)

pour $\mathcal{J}_3$:

$$\begin{aligned}
\forall j \in I_2, \forall k \in I_Z : \mathcal{J}_3(k,\tilde{j}) &= \frac{\partial}{\partial \sigma_j} \left(\frac{1}{\sigma^2_{\bar{Z}_{k,2}}} (X^{\bar{Z}_{k,1}})^t (X^{\bar{Z}_{k,2}} - X^{\bar{Z}_{k,1}} B^{\bar{Z}_{k,2}}_{\bar{Z}_{k,1}}) + A_{\bar{Z}_{k,2}} \lambda^t_{\bar{Z}_{k,1}} \right) && (93)\\
&= 0 \text{ si } \bar{Z}_{k,2} \neq j && (94)
\end{aligned}$$

Donc $\mathcal{J}_3$ est diagonale par blocs (verticaux). Pour un j donné tel que $\bar{Z}_{k,2} = j$, on a

$$\forall k \in I_Z : \mathcal{J}_3(k,\tilde{j}) = -\frac{2}{\sigma_j^3} (X^{\bar{Z}_{k,1}})^t (X^j - X^{\bar{Z}_{k,1}} B^j_{\bar{Z}_{k,1}}) \tag{95}$$

Où $\tilde{j}$ est la position de j dans I_2. (sans perte de généralité, on peut supposer l'ordonnancement des indices mais ça ne simplifierait pas tant que ça (ajout de p_1))

Pour $\mathcal{J}_2$:

$$\begin{aligned}
\forall i \in I_1, \forall k \in I_Z : \mathcal{J}_2(k,i) &= \frac{\partial}{\partial \lambda_i}\left(\frac{1}{\sigma^2_{\bar{Z}_{k,2}}}(X^{\bar{Z}_{k,1}})^t(X^{\bar{Z}_{k,2}} - X^{\bar{Z}_{k,1}}B^{\bar{Z}_{k,2}}_{\bar{Z}_{k,1}}) + A_{\bar{Z}_{k,2}}\lambda^t_{\bar{Z}_{k,1}}\right) && (96)\\
&= \frac{\partial}{\partial \lambda_i}\left(A_{\bar{Z}_{k,2}}\lambda_{\bar{Z}_{k,1}}\right) && (97)\\
&= 0 \text{ si } \bar{Z}_{k,1} \neq i && (98)\\
&= A_{\bar{Z}_{k,2}} \text{ sinon} && (99)
\end{aligned}$$

Enfin, pour $\mathcal{J}_1$:

$$\begin{aligned}
\forall (i,j) \in I_Z^2 : \mathcal{J}_1(i,j) &= \frac{\partial}{\partial B^{\bar{Z}_{j,2}}_{\bar{Z}_{j,1}}}\left(\frac{1}{\sigma^2_{\bar{Z}_{i,2}}}(X^{\bar{Z}_{i,1}})^t(X^{\bar{Z}_{i,2}} - X^{\bar{Z}_{i,1}}B^{\bar{Z}_{i,2}}_{\bar{Z}_{i,1}}) + A_{\bar{Z}_{i,2}}\lambda^t_{\bar{Z}_{i,1}}\right) && (100)\\
&= \frac{\partial}{\partial B^{\bar{Z}_{j,2}}_{\bar{Z}_{j,1}}}\left(-\frac{1}{\sigma^2_{\bar{Z}_{i,2}}}(X^{\bar{Z}_{i,1}})^t X^{\bar{Z}_{i,1}}B^{\bar{Z}_{i,2}}_{\bar{Z}_{i,1}}\right) && (101)\\
&= 0 \text{ si } i \neq j && (102)\\
&= -\frac{1}{\sigma^2_{\bar{Z}_{i,2}}}(X^{\bar{Z}_{i,1}})^t X^{\bar{Z}_{i,1}} \text{ sinon} && (103)
\end{aligned}$$

Donc $\mathcal{J}_1$ est diagonale (au sens strict, pas seulement par blocs).

On va résoudre le système par récurrence, avec comme valeur initiale pour les paramètres recherchés les valeurs actuelles de $\hat{B}$ et $\hat{\Sigma}_{X^{I_2}}$, et pour λ une valeur arbitraire (par exemple vecteur de 1). le terme général de la suite est le suivant :

$$\theta^{(n+1)} = \theta^{(n)} - \mathcal{J}^{-1}(\theta^{(n)})F(\theta^{(n)}) \tag{104}$$

où F est le système (vecteur) des dérivées partielles à annuler.$\theta = (B, \lambda, \sigma)$ avec B en lecture par colonne. Pour la toute première étape B, on utilise comme valeur initiale le B des moindres carrés qui est normalement déjà assez bon donc l'algorithme devrait converger assez vite.

5 Modèle génératif version p_2

On suppose qu'il existe $p2 < p$ sous-régressions qui forment une structure au sein d'un ensemble de p variables X. On définit la structure $S = (p_2, I_2, p_1, I_1)$ comme suit :

$$
\begin{aligned}
I_2 &= (I_2^1, \ldots, I_2^{p_2}) \text{ vecteur des indices des variables dépendantes} && (105)\\
I_1 &= (I_1^1, \ldots, I_1^{p_2}) \text{ vecteur des vecteurs des indices des variables explicatives avec} && (106)\\
I_1^j &\quad \text{ vecteur des indices des variables qui expliquent la variable } I_2^j && (107)\\
p_1 &= (p_1^1, \ldots, p_1^{p_2}) \text{ où } p_1^j = \sharp I_1^j && (108)
\end{aligned}
$$

On suppose que $I_1 \cap I_2 = \emptyset$, *i.e.* que les variables dépendantes n'expliquent aucune autre variable de X. On note I_2^c l'ensemble des variables qui ne sont pas à gauche. On a alors comme modèle génératif :

$$
\begin{aligned}
Y_{|X,S} = Y_{|X} &= XA + \varepsilon_Y = X^{I_2^c} A_{I_2^c} + X^{I_2} A_{I_2} + \varepsilon_Y && (109)\\
\forall j \in I_2 :\ X^j_{|X^{I_1^j},S} &= X^{I_1^j} B^j_{I_1^j} + \varepsilon_j && (110)\\
\forall j \notin I_2 :\ X^j &= f(\theta_j) \text{ loi quelconque} && (111)
\end{aligned}
$$

6 BIC*

On a p_2 sous régression et on note $p_1^j = \sharp I_1^j$ le nombre de variables à droite dans la sous-régression $j \in I_2$. On note $I_1 = (I_1^1, \ldots, I_1^{p_2})$ indices des variables à droite dans chaque régression. Ce vecteur contient donc éventuellement des doublons, mais cela ne remet pas en cause les notations du type

$i \in I_1$. Par contre, le nombre de variables dans l'explicatif est uniquement défini par $(p - p_2)$. On note (nouvelle notation) $p_1 = (p_1^1, \ldots, p_1^{p_2})$

$$P(Z) = P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2) \tag{112}$$

$$P(I_1|p_1, I_2, p_2) = \prod_{j=1}^{p_2} P(I_1^j|p_1^j, I_2, p_2) \tag{113}$$

$$P(I_1^j|p_1^j, I_2, p_2) = \left(\begin{array}{c} p - p_2 \\ p_1^j \end{array} \right)^{-1} = \frac{p_1^j!(p - p_2 - p_1^j)!}{(p - p_2)!} \tag{114}$$

$$P(p_1|I_2, p_2) = \prod_{j=1}^{p_2} P(p_1^j|I_2, p_2) \tag{115}$$

$$P(p_1^j|I_2, p_2) = \frac{1}{p - p_2} \text{ au lieu de } \left(\begin{array}{c} p - p_2 \\ p_1^j \end{array} \right) \frac{1}{2^{(p-p_2)} - 1} \tag{116}$$

$$P(I_2|p_2) = \left(\begin{array}{c} p \\ p_2 \end{array} \right)^{-1} = \frac{p_2!(p - p_2)!}{p!} \tag{117}$$

$$P(p_2) = \frac{1}{p_2} \tag{118}$$

$$P(Z) = \left(\prod_{j=1}^{p_2} \left(\begin{array}{c} p - p_2 \\ p_1^j \end{array} \right)^{-1} \right) \left(\frac{1}{p - p_2} \right)^{p_2} \frac{p_2!(p - p_2)!}{p!} \frac{1}{p_2} \tag{119}$$

$$\ln P(Z) = -\sum_{j=1}^{p_2} \ln \left(\begin{array}{c} p - p_2 \\ p_1^j \end{array} \right) - p_2 \ln(p - p_2) - \ln \left(\begin{array}{c} p \\ p_2 \end{array} \right) - \ln(p_2) \tag{120}$$

Il y a plus de modèles réalisables à 2 sous-régressions que de modèles à 1 sous-régression donc donner à $P(p_2)$ une loi a priori uniforme revient (dans la limite où $p_2 < \frac{p}{2}$) à réduire la probabilité des modèles complexes . Le nombre de modèles à p_2 sous-régressions est le nombre de combinaisons de p_2 variables parmi p multiplié par le produit des nombres de modèles possibles pour chaque sous-régression. On définit alors

$$nb(p_2) = \left(\begin{array}{c} p \\ p_2 \end{array} \right) (2^{(p-p_2)} - 1)^{p_2} \tag{121}$$

$$\sharp\mathcal{S} = \sum_{k=0}^{p-1} nb(k) \text{ le nombre de structures réalisables} \tag{122}$$

avec $nb(k)$ le nombre de modèles à k sous-régressions. On connaît donc au passage le nombre de matrice carrées binaires nilpotentes de taille p.

BIC^* n'a de sens que pour les valeurs de $p_2 < \frac{p}{2}$, au-delà de quoi la pénalité devient contre-productive. En effet, BIC^* favorise les modèles rares (ceux avec peu de sous-régressions ou ceux avec énormément de sous-régressions).Si on décide d'autoriser $p_2 > \frac{p}{2}$ alors il suffit de revenir au BIC classique, ce qui revient à considérer toutes les structures comme étant équiprobables quel que soit le nombre de sous-régressions associées.