

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 :

$$M_1 : Y|X^{I_1}, X^{I_2}, Z \sim X^{I_1} A_{I_1} + X^{I_2} A_{I_2} + \varepsilon \quad (1)$$

$$M_2 : X^{I_2}|X^{I_1}, Z \sim X^{I_1} B_{I_1}^{I_2} + \varepsilon_{X^{I_2}} \quad (2)$$

$$M_3 : X^{I_1}|Z \sim GM(\alpha) \quad (3)$$

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

$$P_{M,Z}(Y, X^{I_1}, X^{I_2}|\theta) = P_{M,Z}(X^{I_1}|\theta)P_{M,Z}(X^{I_2}|X^{I_1};\theta)P_{M,Z}(Y|X^{I_2}, X^{I_1};\theta) \quad (4)$$

$$= P_{M_3,Z}(X^{I_1}|\alpha)P_{M_1,M_2,Z}(X^{I_2}|X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2})P_{M_1,M_2,Z}(Y|X^{I_2}, X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2}) \quad (5)$$

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

$$P_{M_2,Z}(X^{I_2}|X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2}) = P_{M_2,Z}(X^{I_2}|X^{I_1};\Sigma_{X^{I_2}}, B_{I_1}^{I_2}) \text{ et} \quad (6)$$

$$P_{M_1,Z}(Y|X^{I_2}, X^{I_1};\Sigma, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}, A_{I_1}, A_{I_2}) = P_{M_1,Z}(Y|X^{I_2}, X^{I_1};\Sigma, A_{I_1}, A_{I_2}) \quad (7)$$

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 :

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

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 :

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

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

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

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

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

On passe à la log-vraisemblance :

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

2 estimation

On veut maximiser la vraisemblance sous la contrainte $A_{I_1} + B_{I_1}^{I_2} A_{I_2} = \hat{A}_{I_1}$ où \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 $\bar{\Sigma}^{-1}$

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

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

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

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

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

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

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 :

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

et enfin :

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

On a donc :

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

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 première colonne et de sa ligne k . On a alors

$$|\bar{\Sigma}| = (\sigma_Y^2 + A_{I_2}^t \Sigma_{X^{I_2}} A_{I_2}) |\Sigma_X^{I_2}| + \sum_{k=1}^{p_2} (-1)^k A_{I_2(k)} \sigma_{I_2(k)}^2 |\bar{\Sigma}(-1, -(k+1))| \text{ développement 1 colonne} \quad (21)$$

$$= (\sigma_Y^2 + A_{I_2}^t \Sigma_{X^{I_2}} A_{I_2}) |\Sigma_X^{I_2}| + \sum_{k=1}^{p_2} (-1)^k A_{I_2(k)} \sigma_{I_2(k)}^2 [(-1)^{k+1} A_{I_2(k)} \sigma_{I_2(k)}^2 \frac{1}{\sigma_{I_2(k)}^2} \prod_{j \in I_2} \sigma_j^2] \text{ developpements k colonne} \quad (22)$$

$$= \sigma_Y^2 \prod_{j \in I_2} \sigma_j^2 + \sum_{k=1}^{p_2} A_{I_2(k)}^2 \sigma_{I_2(k)}^2 \prod_{j \in I_2} \sigma_j^2 - \sum_{k=1}^{p_2} A_{I_2(k)}^2 \sigma_{I_2(k)}^2 \prod_{j \in I_2} \sigma_j^2 \quad (23)$$

$$= \sigma_Y^2 \prod_{j \in I_2} \sigma_j^2 \quad (24)$$

On peut donc calculer plus finement la log-vraisemblance :

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

$$= \dots \quad (26)$$

$$= -\frac{1}{2} \left[(p_2 + 1) \ln(2\pi) + \ln(|\bar{\Sigma}|) + (X^{I_2} - X^{I_1} B_{I_1}^{I_2}) \Sigma_{X^{I_2}}^{-1} (X^{I_2} - X^{I_1} B_{I_1}^{I_2})^t + \frac{1}{\sigma_Y^2} (Y - XA)^2 \right] \quad (27)$$

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

$$\max_A f(A) = \max_A L_{M_1, M_2, Z}(\Sigma, A_{I_1}, A_{I_2}, \Sigma_{X^{I_2}}, B_{I_1}^{I_2}; Y, X^{I_2}|X^{I_1}) \quad (28)$$

$$s.c. \quad A_{I_1} + B_{I_1}^{I_2} A_{I_2} - \hat{A}_{I_1} = 0 \quad (29)$$

En notant $\psi(A) = A_{I_1} + B_{I_1}^{I_2} A_{I_2} - \hat{A}_{I_1} = B_{I_1} A - \hat{A}_{I_1}$ 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 :

$$g(A, \lambda) = \nabla f(A, \sigma_Y^2) + \lambda \nabla \psi(A) = 0 \quad (30)$$

pour $i \in I$:

$$\frac{\partial g(A, \lambda, \sigma_Y^2)}{\partial A_i} = \frac{\partial}{\partial A_i} \left(-\frac{1}{2\sigma_Y^2} (Y - XA)^t (Y - XA) + \lambda (B_{I_1} A - \hat{A}_{I_1}) \right) \quad (31)$$

$$= -\frac{1}{2\sigma_Y^2} \left(-2(X^i)^t Y + 2(X^i)^t XA \right) + \lambda B_{I_1}^i \quad (32)$$

$$= \frac{1}{\sigma_Y^2} (X^i)^t (Y - XA) + \lambda B_{I_1}^i \quad (33)$$

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 :

$$\frac{\partial g(A, \lambda, \sigma_Y^2)}{\partial \lambda_i} = \frac{\partial \lambda \psi(A)}{\partial \lambda_i} = (B_i A - \hat{A}_i) = A_i + \sum_{j \in I_2} B_{i,j} A_j - \hat{A}_i \quad (34)$$

$$\frac{\partial g(A, \lambda, \sigma_Y^2)}{\partial \sigma_Y} = -\frac{2}{\sigma^3} \sum_{k=1}^n (Y_k - X_k A)^2 \quad (35)$$

L'annulation de cette dérivée ne dépend pas de σ_Y et donne juste

$$\sum_{k=1}^n (Y_k - X_k A)^2 = 0 \quad (36)$$

qui décrit simplement l'absence de biais de l'estimateur.

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)

$$A_{I_1} + B_{I_1}^{I_2} A_{I_2} - \hat{A}_{I_1} = 0 \quad (37)$$

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

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

ce système s'écrit également :

$$A_{I_1} = \hat{A}_{I_1} - B_{I_1}^{I_2} A_{I_2} \quad (40)$$

$$\frac{1}{\sigma_Y^2} (X^{I_1})^t (Y - X^{I_1} (\hat{A}_{I_1} - B_{I_1}^{I_2} A_{I_2}) - X^{I_2} A_{I_2}) + \lambda^t = 0 \quad (41)$$

$$\frac{1}{\sigma_Y^2} (X^{I_2})^t (Y - X^{I_1} (\hat{A}_{I_1} - B_{I_1}^{I_2} A_{I_2}) - X^{I_2} A_{I_2}) + (B_{I_1}^{I_2})^t \lambda^t = 0 \quad (42)$$

On en déduit :

$$\lambda^t = -\frac{1}{\sigma_Y^2} (X^{I_1})^t (Y - X^{I_1} A_{I_1} - X^{I_2} A_{I_2}) \quad (43)$$

$$\text{d'où } 0 = \sigma_Y^2 \frac{1}{\sigma_Y^2} (X^{I_2})^t (Y - X^{I_1} (\hat{A}_{I_1} - B_{I_1}^{I_2} A_{I_2}) - X^{I_2} A_{I_2}) - \sigma_Y^2 \frac{1}{\sigma_Y^2} (B_{I_1}^{I_2})^t (X^{I_1})^t (Y - X^{I_1} A_{I_1} - X^{I_2} A_{I_2}) \quad (44)$$

$$0 = (X^{I_2} - X^{I_1} B_{I_1}^{I_2})^t (Y - X^{I_1} \hat{A}_{I_1} - (X^{I_2} - X^{I_1} B_{I_1}^{I_2}) A_{I_2}) \quad (45)$$

$$\text{d'où } (X^{I_2} - X^{I_1} B_{I_1}^{I_2})^t (Y - X^{I_1} \hat{A}_{I_1}) = (X^{I_2} - X^{I_1} B_{I_1}^{I_2})^t (X^{I_2} - X^{I_1} B_{I_1}^{I_2}) A_{I_2} \quad (46)$$

$$\text{et enfin } \hat{A}_{I_2} = \left((X^{I_2} - X^{I_1} B_{I_1}^{I_2})^t (X^{I_2} - X^{I_1} B_{I_1}^{I_2}) \right)^{-1} (X^{I_2} - X^{I_1} B_{I_1}^{I_2})^t (Y - X^{I_1} \hat{A}_{I_1}) \quad (47)$$

$$\text{puis } \hat{A}_{I_1} = \hat{A}_{I_1} - B_{I_1}^{I_2} A_{I_2} \quad (48)$$

On remarque que $(X^{I_2} - X^{I_1} B_{I_1}^{I_2}) = \varepsilon_{X^{I_2}}$ résidu des sous-régressions et que $Y - X^{I_1} \hat{A}_{I_1} = \tilde{\varepsilon}$ résidu du modèle explicatif. On se contente donc en fait d'estimer les résidus du modèle explicatif par une seconde régression linéaire.

$$Y - X^{I_1} \hat{A}_{I_1} = \varepsilon_{X^{I_2}} A_{I_2} + \varepsilon_Y \quad (49)$$

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 :

$$\nabla f(B) + \lambda \nabla \psi(B) = 0 \quad (50)$$

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

$$g(B, \lambda, \Sigma_{X^{I_2}}) = \nabla f(B, \Sigma_{X^{I_2}}) + \lambda \nabla \psi(B) = 0 \quad (51)$$

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

$$\frac{\partial g(B, \lambda, \Sigma_{X^{I_2}})}{\partial B_{i,j}} = \frac{\partial}{\partial B_{i,j}} \sum_{k=1}^n \left(-\frac{1}{2} [(X_k^{I_2} - X_k^{I_1} B_{I_1}^{I_2}) \Sigma_{X^{I_2}}^{-1} (X_k^{I_2} - X_k^{I_1} B_{I_1}^{I_2})^t] \right) + \frac{\partial}{\partial B_{i,j}} \left(\lambda (B_{I_1} A - \hat{A}) \right) \quad (52)$$

$$= -\frac{1}{2} \frac{\partial}{\partial B_{i,j}} \left(\sum_{l \in I_2} \frac{1}{\sigma_l^2} (X^l - \sum_{k \in I_1} X^k B_{k,l})^t (X^l - \sum_{k \in I_1} X^k B_{k,l}) \right) + \lambda_i A_j \quad (53)$$

$$= \frac{1}{\sigma_j^2} (X^j - \sum_{k \in I_1} X^k B_{k,j})^t X^i + \lambda_i A_j \quad (54)$$

$$= \frac{1}{\sigma_j^2} (X^j - \sum_{k \in I_1^j} X^k B_{k,j})^t X^i + \lambda_i A_j \quad (55)$$

$$\frac{\partial g(B, \lambda, \Sigma_{X^{I_2}})}{\partial B_{I_1^j}^j} = \frac{1}{\sigma_j^2} (X^{I_1^j})^t (X^j - X^{I_1^j} B_{I_1^j}^j) + A_j \lambda_{I_1^j}^t \quad (56)$$

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

$$\frac{\partial g(B, \lambda, \Sigma_{X^{I_2}})}{\partial \lambda_i} = \frac{\partial \lambda \psi(B)}{\partial \lambda_i} = (B_i A - \hat{A}_i) = A_i + \sum_{j \in I_2} B_{i,j} A_j - \hat{A}_i = 0 \quad (57)$$

$$\frac{\partial g(B, \lambda, \Sigma_{X^{I_2}})}{\partial \sigma_j} = \frac{\partial}{\partial \sigma_j} \left(\sum_{i=1}^n \sum_{k \in I_2} \frac{1}{\sigma_k^2} (X_i^k - X_i^{I_1} B_{I_1}^k)^2 \right) \quad (58)$$

$$= -\frac{2}{\sigma_j^3} \sum_{i=1}^n (X_i^j - X_i^{I_1} B_{I_1}^j)^2 = -\frac{2}{\sigma_j^3} (X^j - X^{I_1} B_{I_1}^j)^t (X^j - X^{I_1} B_{I_1}^j) \quad (59)$$

$$= -\frac{2}{\sigma_j^3} (X^j - X^{I_1^j} B_{I_1^j}^j)^t (X^j - X^{I_1^j} B_{I_1^j}^j) \quad (60)$$

Là encore, l'annulation de la dérivée donne juste l'absence de biais pour les sous-régressions : $(X^j - X^{I_1^j} B_{I_1^j}^j)^t (X^j - X^{I_1^j} B_{I_1^j}^j) = 0$

Et donc on obtient le système :

$$\forall j \in I_2 : \frac{1}{\sigma_j^2} (X^{I_1^j})^t (X^j - X^{I_1^j} B_{I_1^j}^j) + A_j \lambda_{I_1^j}^t = 0 \quad (61)$$

$$A_{I_1} + \sum_{j \in I_2} A_j B_{I_1}^j - \hat{A}_{I_1} = 0 \quad (62)$$

$$\forall j \in I_2 : (X^j - X^{I_1^j} B_{I_1^j}^j)^t (X^j - X^{I_1^j} B_{I_1^j}^j) = 0 \quad (63)$$

On a donc

$$\frac{1}{\sigma_j^2} (X^{I_1^j})^t X^j - \frac{1}{\sigma_j^2} (X^{I_1^j})^t X^{I_1^j} B_{I_1^j}^j + A_j \lambda_{I_1^j}^t = 0 \quad (64)$$

$$(X^{I_1^j})^t X^{I_1^j} B_{I_1^j}^j = \sigma_j^2 \frac{1}{\sigma_j^2} (X^{I_1^j})^t X^j + \sigma_j^2 A_j (\lambda_{I_1^j})^t \quad (65)$$

$$B_{I_1^j}^j = \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \left[(X^{I_1^j})^t X^j + \sigma_j^2 A_j (\lambda_{I_1^j})^t \right] \quad (66)$$

On constate ici que si A_j ou σ_j est nul, alors on revient à l'estimateur des moindres carrés classiques pour la sous-régression concernée. On peut donc continuer en supposant que A_j et σ_j sont strictement non nuls (puisque sinon, on connaît les $B_{i,j}$ correspondant et donc on n'a plus besoin de les chercher).

$$B_{I_1^j}^j = \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \left[(X^{I_1^j})^t X^j + \sigma_j^2 A_j (\lambda_{I_1^j})^t \right] \quad (67)$$

$$B_{I_1^j}^j = \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j + \sigma_j^2 A_j \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (\lambda_{I_1^j})^t \quad (68)$$

Soit J la matrice identité de taille p . On introduit une matrice de passage $J_{I_1}^{I_1^j}$ matrice identité de taille p_1 restreinte en nombre de colonnes. On peut donc écrire la contrainte sous la forme

$$A_{I_1} + \sum_{j \in I_2} J_{I_1}^{I_1^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \left[(X^{I_1^j})^t X^j + \sigma_j^2 A_j (\lambda_{I_1})^t \right] A_j = \hat{A}_{I_1} \quad (69)$$

d'où

$$\sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \left[(X^{I_1^j})^t X^j + \sigma_j^2 A_j (\lambda_{I_1^j})^t \right] A_j = \hat{A}_{I_1} - A_{I_1} \quad (70)$$

$$\sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j A_j + \sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \sigma_j^2 A_j (\lambda_{I_1^j})^t A_j = \hat{A}_{I_1} - A_{I_1} \quad (71)$$

$$\sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \sigma_j^2 A_j^2 (\lambda_{I_1^j})^t = (\hat{A}_{I_1} - A_{I_1}) - \sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j A_j \quad (72)$$

$$\sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \sigma_j^2 A_j^2 J_{I_1^j}^{I_1} \lambda^t = (\hat{A}_{I_1} - A_{I_1}) - \sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j A_j \quad (73)$$

On a donc enfin

$$\lambda^t = \left[\sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} \sigma_j^2 A_j^2 J_{I_1^j}^{I_1} \right]^{-1} \left[(\hat{A}_{I_1} - A_{I_1}) - \sum_{j \in I_2} J_{I_1}^{I_j^j} \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j A_j \right] \quad (74)$$

On en déduit donc, $\forall j \in I_2$:

$$B_{I_1^j}^j = \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j + \sigma_j^2 A_j \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (\lambda_{I_1^j})^t \quad (75)$$

$$= \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j + \sigma_j^2 A_j \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} J_{I_1^j}^{I_1} \lambda^t \quad (76)$$

$$= \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j \quad (77)$$

$$+ \sigma_j^2 A_j \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} J_{I_1^j}^{I_1} \left[\sum_{k \in I_2} J_{I_1}^{I_k^k} \left((X^{I_1^k})^t X^{I_1^k} \right)^{-1} \sigma_k^2 A_k^2 J_{I_1^k}^{I_1} \right]^{-1} \left[(\hat{A}_{I_1} - A_{I_1}) - \sum_{k \in I_2} J_{I_1}^{I_k^k} \left((X^{I_1^k})^t X^{I_1^k} \right)^{-1} (X^{I_1^k})^t X^k A_k \right] \quad (78)$$

On note $\hat{\beta}_k$ le vecteur des paramètres de la sous-régression associée à X^k (à gauche) dans le cadre d'une estimation classique par moindres carrés sous la seule contrainte Z .

$$B_{I_1^j}^j = \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j \quad (79)$$

$$+ \sigma_j^2 A_j \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} J_{I_1^j}^{I_1} \left[\sum_{k \in I_2} J_{I_1}^{I_k^k} \left((X^{I_1^k})^t X^{I_1^k} \right)^{-1} \sigma_k^2 A_k^2 J_{I_1^k}^{I_1} \right]^{-1} \left[(\hat{A}_{I_1} - A_{I_1}) - \sum_{k \in I_2} J_{I_1}^{I_k^k} \left((X^{I_1^k})^t X^{I_1^k} \right)^{-1} (X^{I_1^k})^t X^k A_k \right] \quad (80)$$

$$= \hat{\beta}_j + \sigma_j^2 A_j \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} J_{I_1^j}^{I_1} \left[\sum_{k \in I_2} J_{I_1}^{I_k^k} \left((X^{I_1^k})^t X^{I_1^k} \right)^{-1} \sigma_k^2 A_k^2 J_{I_1^k}^{I_1} \right]^{-1} \left[(\hat{A}_{I_1} - A_{I_1}) - \sum_{k \in I_2} J_{I_1}^{I_k^k} \hat{\beta}_k A_k \right] \quad (81)$$

Dans le cadre d'un modèle naïf (toutes les variables à gauche dépendent de toutes les variables à droite), on a $I_1^j = I_1$ et donc

$$B_{I_1}^j = \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} (X^{I_1^j})^t X^j \quad (82)$$

$$+ \sigma_j^2 A_j \left((X^{I_1^j})^t X^{I_1^j} \right)^{-1} J_{I_1^j}^{I_1} \left[\sum_{k \in I_2} J_{I_1}^{I_k^k} \left((X^{I_1^k})^t X^{I_1^k} \right)^{-1} \sigma_k^2 A_k^2 J_{I_1^k}^{I_1} \right]^{-1} \left[(\hat{A}_{I_1} - A_{I_1}) - \sum_{k \in I_2} J_{I_1}^{I_k^k} \left((X^{I_1^k})^t X^{I_1^k} \right)^{-1} (X^{I_1^k})^t X^k A_k \right] \quad (83)$$

$$= \left((X^{I_1})^t X^{I_1} \right)^{-1} (X^{I_1})^t X^j \quad (84)$$

$$+ \sigma_j^2 A_j \left((X^{I_1})^t X^{I_1} \right)^{-1} \left[\sum_{k \in I_2} \left((X^{I_1})^t X^{I_1} \right)^{-1} \sigma_k^2 A_k^2 \right]^{-1} \left[(\hat{A}_{I_1} - A_{I_1}) - \sum_{k \in I_2} \left((X^{I_1})^t X^{I_1} \right)^{-1} (X^{I_1})^t X^k A_k \right] \quad (85)$$

$$= \hat{\beta}_j + \sigma_j^2 A_j \frac{1}{\sum_{k \in I_2} \sigma_k^2 A_k^2} \left[(\hat{A}_{I_1} - A_{I_1}) - \sum_{k \in I_2} \hat{\beta}_k A_k \right] \quad (86)$$

3 Methode de Newton

On veut résoudre le système non linéaire suivant :

$$f1 : \forall j \in I_2 : \frac{1}{\sigma_j^2} (X^{I_1^j})^t (X^j - X^{I_1^j} B_{I_1^j}^j) + A_j \lambda_{I_1^j}^t = 0 \quad (87)$$

$$f2 : A_{I_1} + \sum_{j \in I_2} A_j B_{I_1}^j - \hat{A}_{I_1} = 0 \quad (88)$$

$$f3 : \forall j \in I_2 : (X^j - X^{I_1^j} B_{I_1^j}^j)^t (X^j - X^{I_1^j} B_{I_1^j}^j) = 0 \quad (89)$$

On commence donc par calculer la jacobienne associée :

$$\mathcal{J} = \begin{pmatrix} \mathcal{J}_1 & \mathcal{J}_2 & \mathcal{J}_3 \\ \mathcal{J}_4 & \mathcal{J}_5 & \mathcal{J}_6 \\ \mathcal{J}_7 & \mathcal{J}_8 & \mathcal{J}_9 \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial B_i^j} & \frac{\partial f_1}{\partial \lambda_i} & \frac{\partial f_1}{\partial \sigma_j} \\ \frac{\partial f_2}{\partial B_i^j} & \frac{\partial f_2}{\partial \lambda_i} & \frac{\partial f_2}{\partial \sigma_j} \\ \frac{\partial f_3}{\partial B_i^j} & \frac{\partial f_3}{\partial \lambda_i} & \frac{\partial f_3}{\partial \sigma_j} \end{pmatrix} \quad (90)$$

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 $I_Z = \{1, \dots, p_Z\}$. 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 $\mathcal{J}_5, \mathcal{J}_6, \mathcal{J}_8, \mathcal{J}_9$ sont nuls

$$\mathcal{J} = \begin{pmatrix} \frac{\partial f_1}{\partial B_i^j} & \frac{\partial f_1}{\partial \lambda_i} & \frac{\partial f_1}{\partial \sigma_j} \\ \frac{\partial f_2}{\partial B_i^j} & 0 & 0 \\ \frac{\partial f_3}{\partial B_i^j} & 0 & 0 \end{pmatrix} \quad (91)$$

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}_7

$$\forall (j, k) \in I_2 \times I_Z, \mathcal{J}_7(j, k) = \frac{\partial}{\partial B_{\bar{Z}_{k,1}}^{\bar{Z}_{k,2}}} \left((X^j - X^{I_1^j} B_{I_1^j}^j)^t (X^j - X^{I_1^j} B_{I_1^j}^j) \right) = 0 \text{ si } \bar{Z}_{k,2} \neq j \quad (92)$$

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

$$\mathcal{J}_7(j, k) = \frac{\partial}{\partial B_{\bar{Z}_{k,1}}^j} \left((X^j - X^{I_1^j} B_{I_1^j}^j)^t (X^j - X^{I_1^j} B_{I_1^j}^j) \right) \quad (93)$$

$$= \frac{\partial}{\partial B_{\bar{Z}_{k,1}}^j} \left((X^j)^t X^j - 2(X^j)^t X^{I_1^j} B_{I_1^j}^j + (X^{I_1^j} B_{I_1^j}^j)^t X^{I_1^j} B_{I_1^j}^j \right) \quad (94)$$

$$= \frac{\partial}{\partial B_{\bar{Z}_{k,1}}^j} \left(-2(X^j)^t \sum_{i \in I_1^j} X^i B_i^j + \sum_{l=1}^n \sum_{i \in I_1^j} (X_l^i B_i^j)^2 \right) \quad (95)$$

$$= -2(X^j)^t X^{\bar{Z}_{k,1}} + 2B_{\bar{Z}_{k,1}}^j (X^{\bar{Z}_{k,1}})^t X^{\bar{Z}_{k,1}} \quad (96)$$

$$= -2(X^j - X^{\bar{Z}_{k,1}} B_{\bar{Z}_{k,1}}^j)^t X^{\bar{Z}_{k,1}} \quad (97)$$

On passe à \mathcal{J}_4 :

$$\forall (i, k) \in I_1 \times I_Z, \mathcal{J}_4(\tilde{i}, k) = \frac{\partial}{\partial B_{\bar{Z}_{k,1}}^{\bar{Z}_{k,2}}} \left(A_i + \sum_{j \in I_2} A_j B_i^j - \hat{A}_i \right) = 0 \text{ si } \bar{Z}_{k,1} \neq i \quad (98)$$

$$= A_{\bar{Z}_{k,2}} \text{ si } \bar{Z}_{k,1} = i \quad (99)$$

où \tilde{i} est la position de i dans I_1 .

pour \mathcal{J}_3 :

$$\forall j \in I_2, \forall k \in I_Z : \mathcal{J}_3(k, \tilde{j}) = \frac{\partial}{\partial \sigma_j} \left(\frac{1}{\sigma_{\bar{Z}_{k,2}}^2} (X^{\bar{Z}_{k,1}})^t (X^{\bar{Z}_{k,2}} - X^{\bar{Z}_{k,1}} B_{\bar{Z}_{k,1}}^{\bar{Z}_{k,2}}) + A_{\bar{Z}_{k,2}} \lambda_{\bar{Z}_{k,1}}^t \right) \quad (100)$$

$$= 0 \text{ si } \bar{Z}_{k,2} \neq j \quad (101)$$

$$= \text{ si } \bar{Z}_{k,2} = j \quad (102)$$

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

$$\forall k \in I_Z : \mathcal{J}_3(k, \tilde{j}) = -\frac{2}{\sigma_j^3} (X^{\bar{Z}_{k,1}})^t (X^j - X^{\bar{Z}_{k,1}} B_{\bar{Z}_{k,1}}^j) \quad (103)$$

Pour \mathcal{J}_2 :

$$\forall i \in I_1, \forall k \in I_Z : \mathcal{J}_2(k, i) = \frac{\partial}{\partial \lambda_i} \left(\frac{1}{\sigma_{\bar{Z}_{k,2}}^2} (X^{\bar{Z}_{k,1}})^t (X^{\bar{Z}_{k,2}} - X^{\bar{Z}_{k,1}} B_{\bar{Z}_{k,1}}^{\bar{Z}_{k,2}}) + A_{\bar{Z}_{k,2}} \lambda_{\bar{Z}_{k,1}}^t \right) \quad (104)$$

$$= \frac{\partial}{\partial \lambda_i} (A_{\bar{Z}_{k,2}} \lambda_{\bar{Z}_{k,1}}) \quad (105)$$

$$= 0 \text{ si } \bar{Z}_{k,1} \neq i \quad (106)$$

$$= A_{\bar{Z}_{k,2}} \text{ sinon} \quad (107)$$

Enfin, pour \mathcal{J}_1 :

$$\forall (i, j) \in I_Z^2 : \mathcal{J}_1(i, j) = \frac{\partial}{\partial B_{\bar{Z}_{j,1}}^{\bar{Z}_{j,2}}} \left(\frac{1}{\sigma_{\bar{Z}_{i,2}}^2} (X^{\bar{Z}_{i,1}})^t (X^{\bar{Z}_{i,2}} - X^{\bar{Z}_{i,1}} B_{\bar{Z}_{i,1}}^{\bar{Z}_{i,2}}) + A_{\bar{Z}_{i,2}} \lambda_{\bar{Z}_{i,1}}^t \right) \quad (108)$$

$$= \frac{\partial}{\partial B_{\bar{Z}_{j,1}}^{\bar{Z}_{j,2}}} \left(-\frac{1}{\sigma_{\bar{Z}_{i,2}}^2} (X^{\bar{Z}_{i,1}})^t X^{\bar{Z}_{i,1}} B_{\bar{Z}_{i,1}}^{\bar{Z}_{i,2}} \right) \quad (109)$$

$$= 0 \text{ si } i \neq j \quad (110)$$

$$= -\frac{1}{\sigma_{\bar{Z}_{i,2}}^2} (X^{\bar{Z}_{i,1}})^t X^{\bar{Z}_{i,1}} \text{ sinon} \quad (111)$$

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 :

$$\theta^{(n+1)} = \theta^{(n)} - \mathcal{J}^{-1} F \quad (112)$$

où F est le système (vecteur) des dérivées partielles à annuler.