Section 4

Modèles additifs généralisés (mixtes)

Séance du 5 février 2027
Lectures : Wu, § 2.5 (et § 2.6 pour l'approximation de Laplace) ; Wood (2017), ch. 4 à 6 ;
complément H26 « Modèles additifs généralisés (mixtes) » (A. Bureau)

Julien Miron – notes rédigées à partir du matériel de Lajmi Lakhal-Chaieb et Alexandre Bureau

Objectifs de la séance

À la fin de la séance, vous devriez être en mesure de :

1. représenter une fonction lisse inconnue d'une covariable continue (le temps, l'âge, une dose) à l'aide d'une base de splines, et choisir la dimension de cette base ;
2. écrire la pénalité de rugosité $\int g''(t)^2 \, \mathrm{d}t$ comme une forme quadratique $\boldsymbol{\gamma}^\top S \boldsymbol{\gamma}$ et expliquer le rôle du paramètre de lissage λ ;
3. dériver l'algorithme des moindres carrés itérativement repondérés et pénalisés (P-IRLS) ;
4. comparer deux façons de choisir λ : la validation croisée généralisée et l'estimation de λ comme une composante de variance (ML ou REML) dans la représentation du GAM comme GLMM ;
5. interpréter les degrés de liberté effectifs et les intervalles de confiance « bayésiens » de mgcv ;
6. ajouter des effets aléatoires de sujet (GAMM) et connaître les limites des outils R pour le faire ;
7. reconnaître quand une spline linéaire par morceaux (paramétrique) est une meilleure option.

Cette séance prolonge la séance 3 (modèles mixtes et GLMM) : on y verra qu'un modèle additif pénalisé *est* un modèle mixte, ce qui permet de réutiliser toute la machinerie (vraisemblance marginale, REML, approximation de Laplace) vue la semaine dernière. La notation suit Wu (2010, § 2.5) pour le modèle et Wood (2017) pour l'estimation, comme dans le document complémentaire de l'hiver 2026.

1 Mise en contexte

Dans une étude longitudinale de l'humeur dépressive chez des personnes qui consomment des drogues, on mesure à chaque visite une réponse binaire Y_{ij} (humeur dépressive oui/non) chez le sujet i au temps t_{ij} (en mois depuis le recrutement), ainsi que des covariables comme le statut de consommation.

Rien ne garantit que la log-cote de dépression évolue linéairement dans le temps : l'effet d'une intervention peut s'installer pendant quelques mois puis plafonner, ou une remontée peut survenir à la fin du suivi. Le devoir 1 porte sur un problème de ce type ; nous travaillerons ici avec des données simulées de structure semblable.

Plusieurs options s'offrent à l'analyste pour représenter l'effet du temps :

— **un polynôme** $\beta_1 t + \beta_2 t^2 + \beta_3 t^3$: simple, mais global (une observation aberrante à $t = 0$ influence la courbe à $t = 36$) et instable aux extrémités ;

— **le temps catégoriel** (un paramètre par visite) : flexible, mais impossible avec des temps de visite irréguliers et coûteux en paramètres ;

— **une spline linéaire par morceaux** avec un nœud fixé a priori (p. ex. un changement de pente à 18 mois) : paramétrique, facile à interpréter, mais exige de connaître l'emplacement du nœud ;

— **une spline pénalisée** : la forme de la courbe est apprise des données, et un paramètre de lissage contrôle le compromis entre ajustement et régularité. C'est l'objet de la séance.

Le modèle additif généralisé (GAM, *generalized additive model*) pour des données indépendantes s'écrit (Wu, éq. 2.52, avec une seule fonction lisse pour simplifier)

$$h\{\mathrm{E}(Y_i)\} = \mathbf{x}_i^\top \boldsymbol{\beta} + g(t_i), \qquad i = 1, \ldots, n, \tag{1}$$

où h est la fonction de lien, Y_i suit une loi de la famille exponentielle et g est une fonction lisse inconnue. Plus généralement, on peut avoir $\sum_{k=1}^{p} g_k(t_{ki})$: c'est le caractère *additif*. Comme une constante peut passer de $\mathbf{x}_i^\top \boldsymbol{\beta}$ à g, on impose une contrainte d'identifiabilité, habituellement $\sum_i g(t_i) = 0$.

2 Représenter une fonction lisse : bases de splines

L'idée centrale est de ramener l'estimation d'une fonction à celle d'un vecteur de coefficients : on choisit des fonctions de base connues $\psi_0, \ldots, \psi_r$ et l'on écrit

$$g(t) = \sum_{l=0}^{r} \gamma_l \psi_l(t) = \boldsymbol{\gamma}^\top \boldsymbol{\psi}(t), \qquad \boldsymbol{\psi}(t) = \{\psi_0(t), \ldots, \psi_r(t)\}^\top. \tag{2}$$

Une fois la base choisie, le modèle (1) devient un GLM ordinaire dont la matrice de schéma contient les colonnes $\boldsymbol{\psi}(t_i)^\top$; on note Ψ la matrice $n \times (r+1)$ de lignes $\boldsymbol{\psi}(t_i)^\top$. Toute la difficulté consiste à choisir une base flexible mais numériquement stable, puis à éviter le surajustement.

2.1 Splines cubiques et base des puissances tronquées

Définition 2.1. Soit des nœuds $\kappa_1 < \cdots < \kappa_K$ dans l'intervalle $[a,b]$. Une *spline cubique* de nœuds $\kappa_1, \ldots, \kappa_K$ est une fonction qui est un polynôme de degré 3 sur chaque intervalle $[\kappa_k, \kappa_{k+1}]$ et qui est deux fois continûment dérivable sur $[a,b]$. Une spline cubique est dite *naturelle* si elle est linéaire hors de $[\kappa_1, \kappa_K]$.

L'espace des splines cubiques à K nœuds intérieurs est un espace vectoriel de dimension $K + 4$: quatre coefficients par intervalle ($4(K+1)$), moins trois contraintes de continuité (de la fonction, de

la dérivée première et de la dérivée seconde) à chacun des K nœuds. Une base explicite est la *base des puissances tronquées*

$$1,\, t,\, t^2,\, t^3,\, (t-\kappa_1)_+^3,\ \ldots,\ (t-\kappa_K)_+^3, \qquad (u)_+ = \max(0,u).$$

Chaque fonction $(t-\kappa_k)_+^3$ ajoute un saut de la dérivée troisième en κ_k sans briser la continuité des dérivées d'ordre inférieur. Cette base est commode pour raisonner mais mauvaise pour calculer : ses colonnes sont très corrélées et la matrice $\Psi^\top\Psi$ est mal conditionnée.

2.2 B-splines

Les B-splines engendrent le même espace avec des fonctions à support local. Pour des nœuds étendus $\tau_1 \le \tau_2 \le \ldots$ (on ajoute des nœuds de part et d'autre de $[a,b]$), on définit par récurrence (formule de Cox–de Boor)

$$B_{l,0}(t) = \mathbf{1}\{\tau_l \le t < \tau_{l+1}\}, \qquad B_{l,m}(t) = \frac{t-\tau_l}{\tau_{l+m}-\tau_l}\, B_{l,m-1}(t) + \frac{\tau_{l+m+1}-t}{\tau_{l+m+1}-\tau_{l+1}}\, B_{l+1,m-1}(t).$$

Les B-splines cubiques ($m=3$) ont les propriétés suivantes : (i) $B_{l,3}(t) \ge 0$; (ii) $B_{l,3}$ est nulle hors de $[\tau_l,\tau_{l+4}]$ (support de quatre intervalles) ; (iii) $\sum_l B_{l,3}(t) = 1$ sur $[a,b]$ (partition de l'unité). La propriété (ii) rend Ψ creuse et bien conditionnée ; la propriété (iii) implique que les constantes sont dans l'espace engendré, ce qui exige la contrainte d'identifiabilité mentionnée plus haut.

2.3 Splines cubiques de régression

La base de `s(..., bs = "cr")` dans mgcv paramètre la spline cubique naturelle par ses *valeurs aux nœuds* : $\gamma_l = g(\kappa_l)$ (Wood, 2017, § 5.3.1). Les fonctions de base sont alors des splines cardinales, qui valent 1 à un nœud et 0 aux autres. L'interprétation des coefficients est immédiate et les nœuds sont placés aux quantiles de la covariable. La base par défaut de `s()`, `bs = "tp"` (spline de régression en plaque mince), est construite autrement (troncature d'une décomposition propre) mais joue le même rôle. La figure 1 illustre deux bases à huit fonctions.

2.4 La dimension de la base

Dans une approche par splines *de régression* non pénalisées, le nombre de nœuds est le seul paramètre de lissage : trop peu de nœuds donne une courbe biaisée, trop de nœuds une courbe qui suit le bruit, et le choix discret du nombre et de l'emplacement des nœuds est délicat. L'approche *pénalisée* découple les deux rôles :

— la dimension $k = r+1$ de la base fixe la flexibilité *maximale* ; on la choisit assez grande pour contenir la vraie fonction à une bonne approximation près (souvent k entre 10 et 20) ;

— un terme de pénalité, réglé par λ, détermine la flexibilité *effective*.

Le résultat dépend alors peu de k pourvu que k soit assez grand. En pratique, on vérifie a posteriori que les degrés de liberté effectifs (section 8) restent nettement sous $k-1$; `gam.check()` fournit un test heuristique en ce sens. Si ce n'est pas le cas, on augmente k.

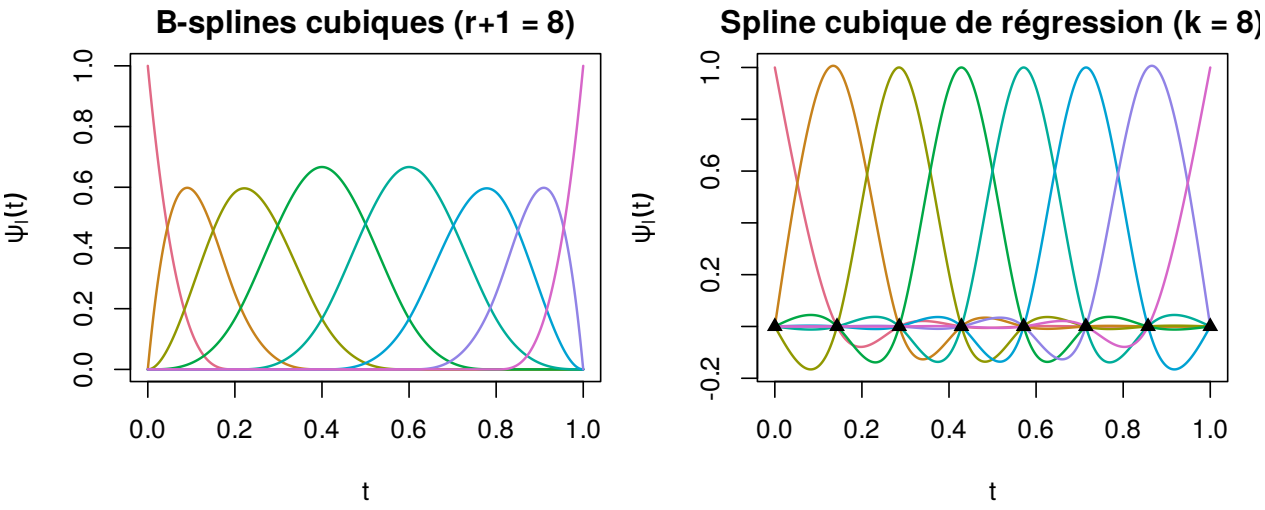


FIGURE 1 – Gauche : huit B-splines cubiques sur [0,1]. Droite : base de spline cubique de régression (`bs = "cr"`, $k = 8$) ; chaque fonction vaut 1 à son nœud (triangles) et 0 aux autres.

À retenir

Une base de splines transforme l'estimation d'une fonction en un problème de régression sur $r+1$ coefficients $\boldsymbol{\gamma}$. Les B-splines et les splines cubiques de régression sont des bases locales et numériquement stables. Dans l'approche pénalisée, k fixe un plafond de flexibilité ; c'est λ qui contrôle la régularité effective.

3 La pénalité de rugosité

On mesure la rugosité d'une fonction deux fois dérivable par l'intégrale du carré de sa courbure approximative, $J(g) = \int \{g''(t)\}^2 \, \mathrm{d}t$. Une droite a une rugosité nulle ; une fonction qui oscille beaucoup a une rugosité élevée.

Proposition 3.1. *Si $g(t) = \boldsymbol{\gamma}^\top \boldsymbol{\psi}(t)$ avec des ψ_l deux fois dérivables, alors*

$$\int \{g''(t)\}^2 \, \mathrm{d}t = \boldsymbol{\gamma}^\top S \boldsymbol{\gamma}, \qquad S_{ll'} = \int \psi_l''(t)\, \psi_{l'}''(t) \, \mathrm{d}t,$$

où S est une matrice $(r+1) \times (r+1)$ symétrique semi-définie positive. Son noyau contient tout $\boldsymbol{\gamma}$ tel que $\boldsymbol{\gamma}^\top \boldsymbol{\psi}$ est une fonction affine de t.

Démonstration. Par linéarité de la dérivation, $g''(t) = \sum_l \gamma_l \psi_l''(t) = \boldsymbol{\gamma}^\top \boldsymbol{\psi}''(t)$. Donc $\{g''(t)\}^2 = \boldsymbol{\gamma}^\top \boldsymbol{\psi}''(t) \boldsymbol{\psi}''(t)^\top \boldsymbol{\gamma}$ et, en intégrant terme à terme, $J(g) = \boldsymbol{\gamma}^\top \{\int \boldsymbol{\psi}''(t) \boldsymbol{\psi}''(t)^\top \, \mathrm{d}t\} \boldsymbol{\gamma} = \boldsymbol{\gamma}^\top S \boldsymbol{\gamma}$. La symétrie est évidente et $\boldsymbol{\gamma}^\top S \boldsymbol{\gamma} = J(g) \geq 0$ pour tout $\boldsymbol{\gamma}$. Enfin, $\boldsymbol{\gamma}^\top S \boldsymbol{\gamma} = 0$ si et seulement si $g'' \equiv 0$ (presque partout, donc partout par continuité), c'est-à-dire si g est affine. □

Pour une base cubique contenant les fonctions affines, le noyau de S est de dimension 2 et $\mathrm{rang}(S) = r - 1$; c'est ce que donne R pour la base `cr` à $k = 8$ (rang 6). Le sous-espace des fonctions affines est donc *non pénalisé* : quelle que soit la valeur de λ, une tendance linéaire est estimée sans contrainte. Quand $\lambda \to \infty$, la fonction estimée tend vers la droite ajustée par maximum de vraisemblance.

Remarque 3.2. Un résultat classique (Green et Silverman, 1994) justifie le choix des splines cubiques : parmi toutes les fonctions deux fois dérivables, celle qui minimise $\sum_i \{y_i - g(t_i)\}^2 + \lambda J(g)$ est une spline cubique naturelle dont les nœuds sont les valeurs distinctes t_i. Les splines de régression pénalisées utilisent une base plus petite ($k \ll n$) qui approxime très bien cette solution à moindre coût. Les P-splines d'Eilers et Marx (1996) remplacent S par une pénalité sur les différences d'ordre 2 des coefficients d'une base de B-splines, $\sum_l (\gamma_{l+1} - 2\gamma_l + \gamma_{l-1})^2$, qui est aussi une forme quadratique $\boldsymbol{\gamma}^\top D^\top D \boldsymbol{\gamma}$.

4 Vraisemblance pénalisée et algorithme P-IRLS

4.1 Le critère

Pour alléger, on absorbe $\mathbf{x}_i^\top \boldsymbol{\beta}$ dans la base (les colonnes de X sont ajoutées à Ψ avec des lignes et colonnes nulles dans S) et l'on considère le paramètre de dispersion ϕ égal à 1, comme pour les lois binomiale et de Poisson. On estime $\boldsymbol{\gamma}$ en maximisant la *log-vraisemblance pénalisée* (Wu, éq. 2.54)

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \tfrac{1}{2}\,\boldsymbol{\gamma}^\top S_\lambda \boldsymbol{\gamma}, \qquad S_\lambda = \lambda S, \tag{3}$$

où $l(\boldsymbol{\gamma}) = \sum_i \log f(y_i \mid t_i, \boldsymbol{\gamma})$ est la log-vraisemblance du GLM qui traite $\boldsymbol{\gamma}$ comme un paramètre fixe. Le facteur $\frac{1}{2}$ est une convention (il simplifie les dérivées).

— $\lambda \to 0$: on retrouve le GLM non pénalisé sur la base complète (risque de surajustement) ;

— $\lambda \to \infty$: $\boldsymbol{\gamma}$ est contraint au noyau de S (fonction affine).

Cas gaussien. Si $Y_i \sim \mathcal{N}(\boldsymbol{\psi}(t_i)^\top \boldsymbol{\gamma}, \sigma^2)$, alors $l(\boldsymbol{\gamma}) = -\|\mathbf{y} - \Psi\boldsymbol{\gamma}\|^2/(2\sigma^2) + \text{cte}$ et maximiser (3) revient à minimiser

$$\|\mathbf{y} - \Psi\boldsymbol{\gamma}\|^2 + \lambda^* \boldsymbol{\gamma}^\top S \boldsymbol{\gamma}, \qquad \lambda^* = \sigma^2 \lambda,$$

dont la solution explicite est $\hat{\boldsymbol{\gamma}} = (\Psi^\top\Psi + \lambda^* S)^{-1}\Psi^\top \mathbf{y}$. Le vecteur ajusté $\hat{\mathbf{y}} = A_\lambda \mathbf{y}$ est linéaire en $\mathbf{y}$, avec la *matrice d'influence* $A_\lambda = \Psi(\Psi^\top\Psi + \lambda^* S)^{-1}\Psi^\top$. C'est λ^* que mgcv rapporte sous le nom `sp` ; nous garderons la notation λ du document H26, qui correspond à λ^*/σ^2 dans le cas gaussien.

4.2 Dérivation de P-IRLS

Rappelons les ingrédients d'un GLM avec lien h : $\eta_i = \boldsymbol{\psi}(t_i)^\top \boldsymbol{\gamma}$, $\mu_i = h^{-1}(\eta_i)$, $\mathrm{Var}(Y_i) = V(\mu_i)$. Le score et l'information de Fisher sont

$$\frac{\partial l}{\partial \boldsymbol{\gamma}} = \Psi^\top W G (\mathbf{y} - \boldsymbol{\mu}), \qquad \mathcal{I}(\boldsymbol{\gamma}) = \Psi^\top W \Psi,$$

où $G = \mathrm{diag}\{h'(\mu_i)\}$ et $W = \mathrm{diag}\{w_i\}$, $w_i = [V(\mu_i)\, h'(\mu_i)^2]^{-1}$.

Proposition 4.1 (P-IRLS)**.** *Une itération de Fisher appliquée à* (3) *à partir de* $\boldsymbol{\gamma}^{[m]}$ *donne*

$$\boldsymbol{\gamma}^{[m+1]} = \left(\Psi^\top W \Psi + S_\lambda\right)^{-1} \Psi^\top W \mathbf{z}, \qquad z_i = \eta_i + h'(\mu_i)(y_i - \mu_i),$$

où W, $\boldsymbol{\eta}$ *et* $\boldsymbol{\mu}$ *sont évalués en* $\boldsymbol{\gamma}^{[m]}$*. Autrement dit, chaque itération résout le problème de moindres carrés pondérés et pénalisés* $\min_\gamma \ \|W^{1/2}(\mathbf{z} - \Psi\boldsymbol{\gamma})\|^2 + \boldsymbol{\gamma}^\top S_\lambda \boldsymbol{\gamma}$.

Démonstration. Le score pénalisé est $\Psi^\top W G(\mathbf{y}-\boldsymbol{\mu}) - S_\lambda \boldsymbol{\gamma}$ et l'information pénalisée est $\Psi^\top W \Psi + S_\lambda$. Le pas de Fisher s'écrit

$$\boldsymbol{\gamma}^{[m+1]} = \boldsymbol{\gamma}^{[m]} + (\Psi^\top W \Psi + S_\lambda)^{-1}\{\Psi^\top W G(\mathbf{y}-\boldsymbol{\mu}) - S_\lambda \boldsymbol{\gamma}^{[m]}\}.$$

En multipliant par $(\Psi^\top W \Psi + S_\lambda)$: $(\Psi^\top W \Psi + S_\lambda)\boldsymbol{\gamma}^{[m+1]} = \Psi^\top W \Psi \boldsymbol{\gamma}^{[m]} + \Psi^\top W G(\mathbf{y}-\boldsymbol{\mu}) = \Psi^\top W\{\boldsymbol{\eta} + G(\mathbf{y}-\boldsymbol{\mu})\} = \Psi^\top W \mathbf{z}$, puisque $\Psi\boldsymbol{\gamma}^{[m]} = \boldsymbol{\eta}$. Les termes en $S_\lambda \boldsymbol{\gamma}^{[m]}$ se sont annulés. Enfin, l'équation normale du problème pondéré pénalisé est précisément $(\Psi^\top W \Psi + S_\lambda)\boldsymbol{\gamma} = \Psi^\top W \mathbf{z}$. □

Algorithme P-IRLS (pour λ fixé).

1. Initialiser $\hat\mu_i = y_i$ (corrigé pour éviter les bornes, p. ex. $(y_i + 0{,}5)/2$ en binomial) et $\hat\eta_i = h(\hat\mu_i)$.
2. Calculer la pseudo-réponse $z_i = \hat\eta_i + h'(\hat\mu_i)(y_i - \hat\mu_i)$ et les poids $w_i = [V(\hat\mu_i) h'(\hat\mu_i)^2]^{-1}$.
3. Résoudre $\hat{\boldsymbol{\gamma}} = (\Psi^\top W \Psi + S_\lambda)^{-1}\Psi^\top W \mathbf{z}$ (par une décomposition QR stable en pratique), puis $\hat{\boldsymbol{\eta}} = \Psi\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\mu}} = h^{-1}(\hat{\boldsymbol{\eta}})$.
4. Répéter 2–3 jusqu'à convergence de la déviance pénalisée.

À convergence, on définit la matrice d'influence du modèle de travail $A_\lambda = \Psi(\Psi^\top W \Psi + S_\lambda)^{-1}\Psi^\top W$, qui généralise celle du cas gaussien.

5 Choisir λ par validation croisée

5.1 Validation croisée ordinaire et généralisée

Dans le cas gaussien, le critère de validation croisée « un à la fois » (*leave-one-out*) est $V_o(\lambda) = n^{-1}\sum_i \{y_i - \hat g^{[-i]}(t_i)\}^2$, où $\hat g^{[-i]}$ est estimée sans l'observation i. Il semble exiger n ajustements ; ce n'est pas le cas.

Proposition 5.1. *Pour le lisseur linéaire $\hat{\mathbf{y}} = A_\lambda \mathbf{y}$ issu des moindres carrés pénalisés,*

$$y_i - \hat g^{[-i]}(t_i) = \frac{y_i - \hat y_i}{1 - A_{ii}}, \qquad \textit{donc} \qquad V_o(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat y_i)^2}{(1-A_{ii})^2}.$$

Démonstration. Posons $M = \Psi^\top \Psi + \lambda^* S$ et $\boldsymbol{\psi}_i = \boldsymbol{\psi}(t_i)$. Sans l'observation i, la matrice devient $M - \boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top$ et le second membre $\Psi^\top \mathbf{y} - \boldsymbol{\psi}_i y_i$. La formule de Sherman–Morrison donne $(M - \boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top)^{-1} = M^{-1} + M^{-1}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top M^{-1}/(1 - A_{ii})$, avec $A_{ii} = \boldsymbol{\psi}_i^\top M^{-1}\boldsymbol{\psi}_i$. En notant $\hat y_i = \boldsymbol{\psi}_i^\top M^{-1}\Psi^\top \mathbf{y}$,

$$\hat g^{[-i]}(t_i) = \boldsymbol{\psi}_i^\top (M - \boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top)^{-1}(\Psi^\top \mathbf{y} - \boldsymbol{\psi}_i y_i) = \Big(\hat y_i - A_{ii} y_i\Big)\Big(1 + \frac{A_{ii}}{1 - A_{ii}}\Big) = \frac{\hat y_i - A_{ii} y_i}{1 - A_{ii}}.$$

Ainsi $y_i - \hat g^{[-i]}(t_i) = \{y_i(1 - A_{ii}) - \hat y_i + A_{ii} y_i\}/(1 - A_{ii}) = (y_i - \hat y_i)/(1 - A_{ii})$. □

La *validation croisée généralisée* (GCV, Craven et Wahba, 1979) remplace chaque A_{ii} par sa moyenne $\mathrm{tr}(A_\lambda)/n$, ce qui rend le critère invariant par rotation des données et plus stable :

$$V_g(\lambda) = \frac{n\,\|\mathbf{y} - A_\lambda \mathbf{y}\|^2}{\{n - \mathrm{tr}(A_\lambda)\}^2}. \tag{4}$$

Pour un GLM, on applique (4) au modèle de travail de P-IRLS, ou l'on remplace la somme des carrés par la déviance : $V_g = nD(\hat{\boldsymbol{\gamma}})/\{n - \mathrm{tr}(A_\lambda)\}^2$. Pour une loi à dispersion connue (binomiale, Poisson), mgcv utilise par défaut le critère UBRE, une version du C_p de Mallows (proche de l'AIC).

5.2 Limites

Le critère GCV est souvent plat près de son minimum et peut avoir plusieurs minima locaux ; il a tendance à sous-lisser (courbes trop ondulées) une fraction non négligeable du temps, surtout pour de petits échantillons ou des réponses corrélées (Wood, 2011). Surtout, il traite les observations comme indépendantes : avec des mesures répétées, la validation croisée « un à la fois » retire une observation alors que les autres observations du même sujet restent dans l'échantillon, ce qui favorise des courbes qui suivent la corrélation intra-sujet. D'où l'intérêt de l'approche suivante.

6 Le GAM comme modèle mixte : λ comme composante de variance

6.1 Interprétation de la pénalité comme loi a priori

La pénalité $\frac{1}{2}\boldsymbol{\gamma}^\top S_\lambda \boldsymbol{\gamma}$ ressemble au logarithme d'une densité normale : si $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, S_\lambda^-)$, où S_λ^- est un inverse généralisé, alors $\log f(\boldsymbol{\gamma}) = -\frac{1}{2}\boldsymbol{\gamma}^\top S_\lambda \boldsymbol{\gamma} + \text{cte}$ et le maximum de (3) est le mode de la loi a posteriori $f(\boldsymbol{\gamma} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\gamma}) f(\boldsymbol{\gamma})$. Comme S n'est pas de plein rang, cette « loi » est impropre dans les directions du noyau de S (variance infinie, c'est-à-dire aucun rétrécissement). Rendons cette idée précise en séparant les deux sous-espaces.

6.2 Reparamétrisation en effets fixes et aléatoires

Soit la décomposition spectrale $S = U\Lambda U^\top$, avec $U = (U_0, U_+)$ orthogonale, $\Lambda = \mathrm{diag}(\mathbf{0}, \Lambda_+)$, où U_0 engendre le noyau de S (dimension M_0, ici 2) et Λ_+ contient les $M = r + 1 - M_0$ valeurs propres strictement positives. Posons

$$\boldsymbol{\beta}_F = U_0^\top \boldsymbol{\gamma}, \qquad \mathbf{u} = \Lambda_+^{1/2} U_+^\top \boldsymbol{\gamma}, \qquad X_F = \Psi U_0, \qquad Z = \Psi U_+ \Lambda_+^{-1/2}.$$

Alors $\Psi\boldsymbol{\gamma} = \Psi(U_0U_0^\top + U_+U_+^\top)\boldsymbol{\gamma} = X_F\boldsymbol{\beta}_F + Z\mathbf{u}$ et $\boldsymbol{\gamma}^\top S \boldsymbol{\gamma} = \boldsymbol{\gamma}^\top U_+ \Lambda_+ U_+^\top \boldsymbol{\gamma} = \mathbf{u}^\top \mathbf{u}$. Le critère (3) devient

$$l_p = l(\boldsymbol{\beta}_F, \mathbf{u}) - \frac{\lambda}{2}\,\mathbf{u}^\top \mathbf{u}.$$

Proposition 6.1 (Représentation GLMM)**.** *Le maximiseur de* (3) *coïncide avec l'estimateur des effets fixes $\boldsymbol{\beta}_F$ et le prédicteur (mode conditionnel) des effets aléatoires $\mathbf{u}$ dans le GLMM*

$$h\{\mathrm{E}(Y_i \mid \mathbf{u})\} = \mathbf{x}_{F,i}^\top \boldsymbol{\beta}_F + \mathbf{z}_i^\top \mathbf{u}, \qquad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1} I_M),$$

où $\mathbf{x}_{F,i}^\top$ et $\mathbf{z}_i^\top$ sont les lignes de X_F et Z. La partie non pénalisée (fonctions affines, plus les covariables paramétriques) joue le rôle d'effets fixes ; la partie pénalisée joue le rôle d'effets aléatoires i.i.d. de variance $\sigma_u^2 = 1/\lambda$.

Démonstration. Dans ce GLMM, la log-densité conjointe de $(\mathbf{y}, \mathbf{u})$ est $\log f(\mathbf{y} \mid \mathbf{u}) + \log f(\mathbf{u}) = l(\boldsymbol{\beta}_F, \mathbf{u}) - \frac{\lambda}{2}\mathbf{u}^\top\mathbf{u} + \frac{M}{2}\log\lambda + \text{cte}$. Pour λ fixé, la maximiser conjointement en $(\boldsymbol{\beta}_F, \mathbf{u})$ revient à maximiser

l_p ; c'est le critère des « équations du modèle mixte » de Henderson pour le cas gaussien et le critère utilisé par l'approximation de Laplace pour un GLMM (Wu, § 2.6). Comme la reparamétrisation est linéaire et bijective (pour U et Λ fixés), les deux maximiseurs se correspondent. □

Dans le cas gaussien, $\mathbf{u} \sim \mathcal{N}(0,\sigma_u^2 I)$ et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\sigma^2 I)$ donnent $\lambda^* = \sigma^2/\sigma_u^2$: le paramètre de lissage est un *rapport de variances*. Une grande variance des effets aléatoires (courbe très variable) correspond à un petit λ ; une variance nulle à $\lambda = \infty$ (droite). La conséquence pratique est majeure : on peut estimer λ comme n'importe quelle composante de variance, par maximum de vraisemblance (ML) ou maximum de vraisemblance restreinte (REML), au lieu de la validation croisée.

Remarque 6.2. Les « effets aléatoires » $\mathbf{u}$ ne représentent pas ici une variabilité entre unités d'une population : c'est un artifice qui exprime l'hypothèse a priori que la fonction est lisse. On dit parfois que la pénalité est une loi a priori « bayésienne empirique » dont l'hyperparamètre λ est estimé.

7 Estimation de λ par ML et REML

7.1 Vraisemblance marginale et approximation de Laplace

Dans la représentation GLMM, la vraisemblance de λ (et de $\boldsymbol{\beta}_F$) s'obtient en intégrant les effets aléatoires :

$$L(\boldsymbol{\beta}_F,\lambda) = \int f(\mathbf{y} \mid \boldsymbol{\beta}_F,\mathbf{u})\, f_\lambda(\mathbf{u})\, \mathrm{d}\mathbf{u} = \Big(\frac{\lambda}{2\pi}\Big)^{M/2} \int \exp\Big\{ l(\boldsymbol{\beta}_F,\mathbf{u}) - \frac{\lambda}{2}\mathbf{u}^\top \mathbf{u} \Big\}\, \mathrm{d}\mathbf{u}.$$

Sauf dans le cas gaussien, l'intégrale n'a pas de forme explicite. On applique l'approximation de Laplace vue à la séance 3 : on développe l'exposant à l'ordre 2 autour de son maximum $\hat{\mathbf{u}}$ (qui est la solution de P-IRLS pour $\boldsymbol{\beta}_F$ fixé). Le hessien de l'exposant vaut $-(Z^\top W Z + \lambda I)$ (approximation de Fisher), de sorte que

$$\int \exp\{\cdots\}\, \mathrm{d}\mathbf{u} \approx \exp\Big\{ l(\boldsymbol{\beta}_F,\hat{\mathbf{u}}) - \frac{\lambda}{2}\hat{\mathbf{u}}^\top \hat{\mathbf{u}} \Big\} (2\pi)^{M/2}\, |Z^\top W Z + \lambda I|^{-1/2}.$$

Proposition 7.1 (Vraisemblance approximative, ML)**.** *À une constante près,*

$$PL(\lambda) \approx l(\hat{\boldsymbol{\gamma}}) - \frac{1}{2}\hat{\boldsymbol{\gamma}}^\top S_\lambda \hat{\boldsymbol{\gamma}} + \frac{1}{2}\log|S_\lambda|_+ - \frac{1}{2}\log|Z^\top W Z + \lambda I_M|, \tag{5}$$

où $|S_\lambda|_+ = \lambda^M \prod_j \Lambda_{+,jj}$ est le produit des valeurs propres non nulles de S_λ (pseudo-déterminant). On maximise PL en λ, $\hat{\boldsymbol{\gamma}}$ étant l'estimateur pénalisé à ce λ (ce qui maximise aussi en $\boldsymbol{\beta}_F$).

Démonstration. On prend le logarithme de l'approximation de Laplace : $\frac{M}{2}\log\lambda$ provient de la constante de normalisation de $f_\lambda(\mathbf{u})$ et s'écrit $\frac{1}{2}\log|S_\lambda|_+$ à une constante près (le produit des $\Lambda_{+,jj}$ ne dépend pas de λ). Les facteurs $(2\pi)^{\pm M/2}$ se simplifient, et $\hat{\mathbf{u}}^\top\hat{\mathbf{u}}\,\lambda = \hat{\boldsymbol{\gamma}}^\top S_\lambda \hat{\boldsymbol{\gamma}}$ par la reparamétrisation. Enfin, en profilant $\boldsymbol{\beta}_F$ on remplace $(\boldsymbol{\beta}_F,\hat{\mathbf{u}})$ par le maximiseur conjoint, soit $\hat{\boldsymbol{\gamma}}$. □

Le document H26 présente cette expression sous la forme abrégée $PL(\lambda) \approx l(\hat{\boldsymbol{\gamma}}) - \frac{1}{2}\hat{\boldsymbol{\gamma}}^\top S_\lambda \hat{\boldsymbol{\gamma}} + \frac{1}{2}\log|S_\lambda| + K$; il faut garder à l'esprit que le terme de déterminant $-\frac{1}{2}\log|Z^\top W Z + \lambda I|$ dépend de λ et fait partie du critère maximisé : c'est lui qui pénalise une valeur de λ trop petite. Les trois termes ont une interprétation claire :

— $l(\hat{\boldsymbol{\gamma}}) - \frac{1}{2}\hat{\boldsymbol{\gamma}}^\top S_\lambda \hat{\boldsymbol{\gamma}}$: qualité de l'ajustement pénalisé (augmente quand $\lambda \downarrow 0$) ;

— $\frac{1}{2}\log|S_\lambda|_+ - \frac{1}{2}\log|Z^\top W Z + \lambda I| = -\frac{1}{2}\sum_j \log(1 + d_j/\lambda)$ (à une constante près), où les d_j sont les valeurs propres de $Z^\top W Z$: terme de complexité, qui tend vers $-\infty$ quand $\lambda \downarrow 0$. Il joue le même rôle que le dénominateur de GCV.

7.2 Vraisemblance restreinte

Comme dans le modèle linéaire mixte, le ML ne tient pas compte des degrés de liberté perdus en estimant les effets fixes $\boldsymbol{\beta}_F$ et sous-estime les composantes de variance. La vraisemblance restreinte corrige ce biais. Dans le cadre de Laplace, elle s'obtient en intégrant *aussi* les effets fixes contre une loi a priori plate, c'est-à-dire en intégrant $\exp\{l_p(\boldsymbol{\gamma})\}$ par rapport à tout le vecteur $\boldsymbol{\gamma}$ (Wood, 2011). On obtient

$$PL_r(\lambda) \approx l(\hat{\boldsymbol{\gamma}}) - \frac{1}{2}\hat{\boldsymbol{\gamma}}^\top S_\lambda \hat{\boldsymbol{\gamma}} + \frac{1}{2}\log|S_\lambda|_+ - \frac{1}{2}\log|\Psi^\top W \Psi + S_\lambda|. \tag{6}$$

Proposition 7.2. *Avec $V = W^{-1} + \lambda^{-1} Z Z^\top$ (matrice de variance marginale du modèle de travail),*

$$PL_r(\lambda) = PL(\lambda) - \frac{1}{2}\log|X_F^\top V^{-1} X_F|.$$

Démonstration. Dans la paramétrisation ($\boldsymbol{\beta}_F$,$\mathbf{u}$) (dont le jacobien est constant en λ), la matrice $\Psi^\top W \Psi + S_\lambda$ devient la matrice par blocs

$$\begin{pmatrix} X_F^\top W X_F & X_F^\top W Z \\ Z^\top W X_F & Z^\top W Z + \lambda I \end{pmatrix}.$$

Par la formule du déterminant d'une matrice par blocs (complément de Schur), son déterminant vaut $|Z^\top W Z + \lambda I| \cdot |X_F^\top \{W - W Z (Z^\top W Z + \lambda I)^{-1} Z^\top W\} X_F|$. Par l'identité de Woodbury, $W - W Z (Z^\top W Z + \lambda I)^{-1} Z^\top W = (W^{-1} + \lambda^{-1} Z Z^\top)^{-1} = V^{-1}$. En comparant (6) et (5), on obtient le résultat. □

On retrouve exactement la correction REML du modèle linéaire mixte, $l_R = l_{ML} - \frac{1}{2}\log|X^\top V^{-1} X|$ (séance 3). Dans le cas gaussien, $W = \sigma^{-2} I$ et (6) est la vraisemblance restreinte exacte ; le document H26 écrit le terme correspondant avec la pseudo-variance résiduelle σ^2 du modèle de travail, $W = \sigma^{-2} I$.

À retenir

— Le GAM pénalisé est un GLMM : partie non pénalisée = effets fixes, partie pénalisée = effets aléatoires de variance $1/\lambda$ (gaussien : $\lambda^* = \sigma^2/\sigma_u^2$).

— On estime λ en maximisant l'approximation de Laplace de la vraisemblance (PL) ou de la vraisemblance restreinte (PL_r). Ces critères ont un terme d'ajustement et un terme de complexité.

— Le ML sous-estime les composantes de variance ; une variance $1/\lambda$ sous-estimée correspond à un λ trop grand, donc à des courbes *plus lisses* qu'avec le REML. C'est une tendance (le biais de $\hat{\sigma}^2$ intervient aussi dans le cas gaussien), pas une règle absolue.

7.3 Optimisation

Les deux critères se maximisent par une imbrication de deux boucles (Wood, 2011) : une boucle externe de Newton sur $\rho = \log \lambda$ (le logarithme garantit $\lambda > 0$ et rend le critère mieux conditionné), et pour chaque valeur proposée de ρ, une boucle interne P-IRLS qui calcule $\hat{\boldsymbol{\gamma}}(\lambda)$ et les poids W. Les expressions PL et PL_r sont évaluées avec les poids de la dernière itération de P-IRLS. L'approche est stable même avec plusieurs paramètres de lissage $\lambda_1, \ldots, \lambda_p$ (un par fonction lisse : $S_\lambda = \sum_k \lambda_k S_k$).

Niveau doctorat : Laplace restreinte et quasi-vraisemblance pénalisée

Considérons le modèle de travail de P-IRLS à convergence : $\mathbf{z} = X_F\boldsymbol{\beta}_F + Z\mathbf{u} + \mathbf{e}$, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, W^{-1})$, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1} I)$, où W est traité comme connu. C'est un modèle linéaire mixte gaussien, dont la log-vraisemblance restreinte exacte est

$$-\tfrac{1}{2}\log|V| - \tfrac{1}{2}\log|X_F^\top V^{-1} X_F| - \tfrac{1}{2}(\mathbf{z} - X_F\hat{\boldsymbol{\beta}}_F)^\top V^{-1}(\mathbf{z} - X_F\hat{\boldsymbol{\beta}}_F).$$

Exercice guidé. (a) Montrez avec l'identité de Woodbury et le lemme du déterminant ($|V| = |W^{-1}|\,|I + \lambda^{-1} Z^\top W Z|$) que ce critère est égal à $PL_r(\lambda)$ dans lequel on remplace $l(\hat{\boldsymbol{\gamma}})$ par son approximation quadratique $-\tfrac{1}{2}\|W^{1/2}(\mathbf{z} - \Psi\hat{\boldsymbol{\gamma}})\|^2$. (b) En déduire que maximiser le REML du modèle de travail, puis mettre à jour W, et itérer, est la méthode de *quasi-vraisemblance pénalisée* (PQL) de Breslow et Clayton (1993). La différence avec la méthode de Laplace « directe » de `gam(method = "REML")` est que la PQL ignore la dépendance de W envers $(\boldsymbol{\beta}_F, \mathbf{u}, \lambda)$ dans le terme de déterminant : c'est l'« approximation de plus concernant les poids » mentionnée dans les notes H26 de la séance 3 (Wu, § 2.6). Cette approximation est médiocre pour des réponses binaires avec peu d'observations par grappe, ce qui explique les biais connus de la PQL (sous-estimation des variances) dans ce contexte.
(c) Dans le cas gaussien, montrez que PL et PL_r sont exacts : l'exposant est alors exactement quadratique en $\mathbf{u}$, et l'approximation de Laplace est exacte.

8 Degrés de liberté effectifs

Dans un modèle linéaire non pénalisé, le nombre de paramètres est $\mathrm{tr}(A)$, la trace de la matrice chapeau. Par analogie, on définit les *degrés de liberté effectifs* (*edf*) d'un modèle pénalisé par

$$\mathrm{edf} = \mathrm{tr}(A_\lambda) = \mathrm{tr}\big\{(\Psi^\top W \Psi + S_\lambda)^{-1} \Psi^\top W \Psi\big\}.$$

La matrice $F = (\Psi^\top W \Psi + S_\lambda)^{-1} \Psi^\top W \Psi$ transforme l'estimateur non pénalisé en estimateur pénalisé ($\hat{\boldsymbol{\gamma}} = F\tilde{\boldsymbol{\gamma}}$) ; ses éléments diagonaux donnent les edf par coefficient, et l'on somme sur les coefficients d'un même terme lisse pour obtenir son edf.

Proposition 8.1. *Dans la paramétrisation de la proposition 6.1, soit $d_1, \ldots, d_M$ les valeurs propres de $Z^\top (W - W X_F (X_F^\top W X_F)^{-1} X_F^\top W) Z$. Alors*

$$\mathrm{edf} = M_0 + \sum_{j=1}^{M} \frac{d_j}{d_j + \lambda} \in [M_0, r + 1].$$

Esquisse. On projette d'abord sur l'orthogonal (au sens de W) de X_F ; la partie non pénalisée contribue M_0 edf. Dans la base propre de la matrice projetée, le problème pénalisé se découple en M

problèmes scalaires où chaque coordonnée non pénalisée est multipliée par $d_j/(d_j+\lambda)$ (rétrécissement de type ridge). La trace est la somme de ces facteurs. □

Chaque direction est donc rétrécie d'autant plus que λ est grand relativement à l'information d_j qu'apportent les données dans cette direction. Les edf varient continûment de M_0 (droite, $\lambda = \infty$) à $r + 1$ (spline non pénalisée, $\lambda = 0$). Dans les sorties de `summary.gam`, la colonne `edf` exclut la constante absorbée par la contrainte d'identifiabilité (l'edf d'un terme lisse va donc de 1 à $k - 1$). La colonne `Ref.df` est une autre définition, $2\,\mathrm{tr}(F) - \mathrm{tr}(FF)$, utilisée comme degrés de liberté de référence pour les tests.

9 Inférence

9.1 Intervalles « bayésiens »

L'interprétation a priori de la pénalité donne, par l'approximation normale de la loi a posteriori,

$$\boldsymbol{\gamma} \mid \mathbf{y} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\gamma}}, V_\gamma), \qquad V_\gamma = (\Psi^\top W \Psi + S_\lambda)^{-1}\phi.$$

Pour une valeur t, $\hat{g}(t) = \boldsymbol{\psi}(t)^\top \hat{\boldsymbol{\gamma}}$ et un intervalle à 95 % est $\hat{g}(t) \pm 1{,}96\,\{\boldsymbol{\psi}(t)^\top V_\gamma \boldsymbol{\psi}(t)\}^{1/2}$; ce sont les bandes tracées par `plot.gam`. Ces intervalles ne sont pas des intervalles de confiance fréquentistes au sens habituel : l'estimateur pénalisé est biaisé (il rétrécit vers une droite), et la variance fréquentiste $V_F = (\Psi^\top W \Psi + S_\lambda)^{-1}\Psi^\top W \Psi (\Psi^\top W \Psi + S_\lambda)^{-1}$ est plus petite que V_γ. La variance bayésienne incorpore en quelque sorte le biais. Nychka (1988) a montré que ces intervalles ont une couverture proche de 95 % *en moyenne sur les valeurs de t* observées (couverture « à travers la fonction »), même si la couverture ponctuelle peut être mauvaise aux endroits où la courbure est forte. Ils ignorent aussi l'incertitude sur λ ; l'option `unconditional = TRUE` de `plot.gam` et `vcov.gam` applique une correction.

9.2 Tests

Pour tester $H_0 : g \equiv 0$, mgcv utilise une statistique de type Wald, $T = \hat{\boldsymbol{\gamma}}_g^\top V_{\gamma,g}^{r-} \hat{\boldsymbol{\gamma}}_g$, où V^{r-} est un pseudo-inverse tronqué au rang (approximativement l'edf), comparée à une loi khi-deux ou F (Wood, 2013). Les valeurs-p sont approximatives : elles tendent à être un peu trop petites quand λ est estimé, et sont plus fiables avec les méthodes ML ou REML qu'avec GCV. Pour les termes `bs = "re"`, le test porte sur une variance nulle, une hypothèse sur le bord de l'espace des paramètres ; la valeur-p rapportée est alors approximative (souvent conservatrice), comme pour le test du rapport de vraisemblance d'une composante de variance vu à la séance 3 (mélange de lois khi-deux).

Pour comparer deux modèles emboîtés (p. ex. effet linéaire contre effet lisse du temps), on peut utiliser un test du rapport de vraisemblance approximatif avec des ajustements par ML (`anova(m0, m1, test = "Chisq")`) ou l'AIC. Il faut ajuster les deux modèles par ML et non par REML lorsque leurs parties fixes diffèrent, pour la même raison que dans les modèles mixtes : les vraisemblances restreintes de modèles à effets fixes différents ne sont pas comparables.

10 Modèles additifs généralisés mixtes (GAMM)

10.1 Le modèle

Avec des données longitudinales Y_{ij}, $j = 1,\ldots,n_i$, on ajoute au prédicteur des effets aléatoires de sujet (Wu, éq. 2.70, ici avec une seule fonction lisse) :

$$h\{\mathrm{E}(Y_{ij} \mid \mathbf{b}_i)\} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{\gamma}^\top \boldsymbol{\psi}(t_{ij}) + \mathbf{z}_{ij}^\top \mathbf{b}_i, \qquad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, D(\boldsymbol{\eta})), \tag{7}$$

avec indépendance conditionnelle des Y_{ij} sachant $\mathbf{b}_i$. Il y a maintenant deux types d'« effets aléatoires » : les coefficients pénalisés de la spline (variance $1/\lambda$) et les effets de sujet $\mathbf{b}_i$ (matrice de covariance $D(\boldsymbol{\eta})$). Rien ne les distingue sur le plan du calcul : on estime *conjointement* λ et $\boldsymbol{\eta}$ par ML ou REML. Il serait inhabituel d'estimer $\boldsymbol{\eta}$ par validation croisée. La version ML de l'approximation de Laplace devient

$$PL(\lambda, \boldsymbol{\eta}) \approx l(\hat{\boldsymbol{\gamma}}, \hat{\mathbf{b}}) - \frac{1}{2}\hat{\boldsymbol{\gamma}}^\top S_\lambda \hat{\boldsymbol{\gamma}} + \frac{1}{2}\log|S_\lambda|_+ - \frac{1}{2}\hat{\mathbf{b}}^\top D^{-1}(\boldsymbol{\eta})\hat{\mathbf{b}} - \frac{1}{2}\log|D(\boldsymbol{\eta})| - \frac{1}{2}\log|\mathcal{H}| + K,$$

où $\hat{\mathbf{b}}$ regroupe les $\hat{\mathbf{b}}_i$, $D(\boldsymbol{\eta})$ désigne ici la matrice bloc-diagonale de tous les sujets et $\mathcal{H}$ est le hessien (au signe près) de l'exposant par rapport à $(\mathbf{u}, \mathbf{b})$. Comme plus haut, le document H26 omet le dernier déterminant dans sa forme abrégée.

10.2 Mise en œuvre en R et mises en garde

Trois outils sont couramment utilisés.

1. `mgcv::gam` avec un terme `s(id, bs = "re")` : l'effet aléatoire d'ordonnée à l'origine est représenté comme une « fonction lisse » dont la base est l'indicatrice de chaque sujet et la pénalité est l'identité. C'est exactement la représentation de la proposition 6.1 à l'envers, et l'estimation se fait par Laplace (`method = "REML"` ou `"ML"`). C'est l'approche utilisée dans le solutionnaire H26.

2. `mgcv::gamm` : ajuste le modèle via `nlme::lme` (cas gaussien) ou par PQL via `MASS::glmmPQL` (cas non gaussien). Pour une réponse binaire, la PQL est biaisée (voir l'encadré précédent) ; on l'évite.

3. `gamm4::gamm4` : représente la spline comme effet aléatoire dans `lme4`, ce qui permet l'approximation de Laplace de `glmer` et une structure d'effets aléatoires générale.

Attention : effets aléatoires de pente

Avec `gam`, un terme `s(id, temps, bs = "re")` ajoute une pente aléatoire *indépendante* de l'ordonnée à l'origine aléatoire : on ne peut pas estimer de corrélation entre les deux, contrairement à `(1 + temps | id)` dans `glmer`. De plus, chaque niveau de `id` ajoute un coefficient au problème pénalisé, si bien que le temps de calcul croît rapidement avec le nombre de sujets. Enfin, une pente aléatoire en t et une fonction lisse $g(t)$ se disputent la même information sur la tendance temporelle. Dans le devoir H26, la partie GAMM ne comprenait qu'une ordonnée à l'origine aléatoire ; la structure avec pentes aléatoires corrélées était étudiée avec une spline linéaire par morceaux dans `glmer` (section suivante). Si l'on veut à la fois une fonction lisse et des pentes aléatoires corrélées, `gamm4` est l'outil approprié.

10.3 Une alternative paramétrique : la spline linéaire par morceaux

Quand le contexte suggère un changement de régime à un moment connu (fin d'une intervention, 18 mois de suivi), on peut remplacer $g(t)$ par

$$g(t) = \beta_1 t + \beta_2 (t - \kappa)_+, \qquad \text{en R : } \texttt{pmax(0, temps - 18)} \text{ pour } \kappa = 18.$$

C'est une spline de degré 1 à un nœud, non pénalisée. La pente est β_1 avant κ et $\beta_1 + \beta_2$ après ; $H_0 : \beta_2 = 0$ teste l'absence de changement de pente. Avantages : paramètres interprétables, tests standards, et possibilité d'effets aléatoires de pente avant et après le nœud avec une matrice D non structurée dans `glmer`. Inconvénients : le nœud doit être choisi a priori (s'il est choisi au vu des données, les tests ne sont plus valides au niveau nominal) et la forme est rigide. Un GAM ajusté au préalable est un bon outil exploratoire pour juger si une telle forme est plausible, comme on le verra dans l'exemple 11.3.

11 Exemples en R

Exemple 11.1 (GAM gaussien : données `mcycle`)**.** Le jeu de données `mcycle` (module MASS) contient l'accélération de la tête (en g) mesurée à 133 instants (en ms) après un impact simulé de motocyclette. La relation est fortement non linéaire. On ajuste une spline de régression pénalisée de base $k = 20$ avec trois méthodes de choix de λ.

```
library(mgcv); library(MASS)
fit_gcv  <- gam(accel ~ s(times, k = 20), data = mcycle, method = "GCV.Cp")
fit_reml <- gam(accel ~ s(times, k = 20), data = mcycle, method = "REML")
fit_ml   <- gam(accel ~ s(times, k = 20), data = mcycle, method = "ML")
summary(fit_reml)
```

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -25.55       1.96  -13.03   <2e-16 ***
Approximate significance of smooth terms:
           edf Ref.df    F p-value
s(times) 12.18  14.59 32.3  <2e-16 ***
R-sq.(adj) =  0.781   Deviance explained = 80.1%
-REML = 617.99  Scale est. = 511.15    n = 133
```

```
sapply(list(GCV = fit_gcv, REML = fit_reml, ML = fit_ml),
       function(f) c(lambda = f$sp, edf = sum(f$edf) - 1))
```

```
                      GCV        REML          ML
lambda.s(times)  0.0022076  0.00128979  0.00133690
edf             10.8980676 12.17616305 12.08954367
```

L'edf de 12,2 est bien inférieur au plafond $k - 1 = 19$: la dimension de base est suffisante. Ici, la GCV choisit une courbe un peu plus lisse que le REML ; le ML est très proche du REML, car un seul paramètre fixe (la pente) est perdu en plus de la constante. La figure 2 (gauche) montre l'effet de λ : quand $\lambda \to \infty$, on retrouve la droite des moindres carrés.

On vérifie la représentation en modèle mixte : la fonction `gam.vcomp` convertit λ en écart type d'effet aléatoire, $\sigma_u = (\hat{\sigma}^2 / \hat{\lambda}^*)^{1/2}$.

```
gam.vcomp(fit_reml, rescale = FALSE)
sqrt(fit_reml$sig2 / fit_reml$sp)
```

```
             std.dev     lower     upper
s(times) 629.52609 413.63149 958.10668
scale     22.60855  19.91681  25.66407
[1] 629.5262
```

(Avec l'option par défaut `rescale = TRUE`, mgcv rapporte l'écart type pour une pénalité renormalisée ; seul le rapport σ^2/σ_u^2 a un sens intrinsèque.)

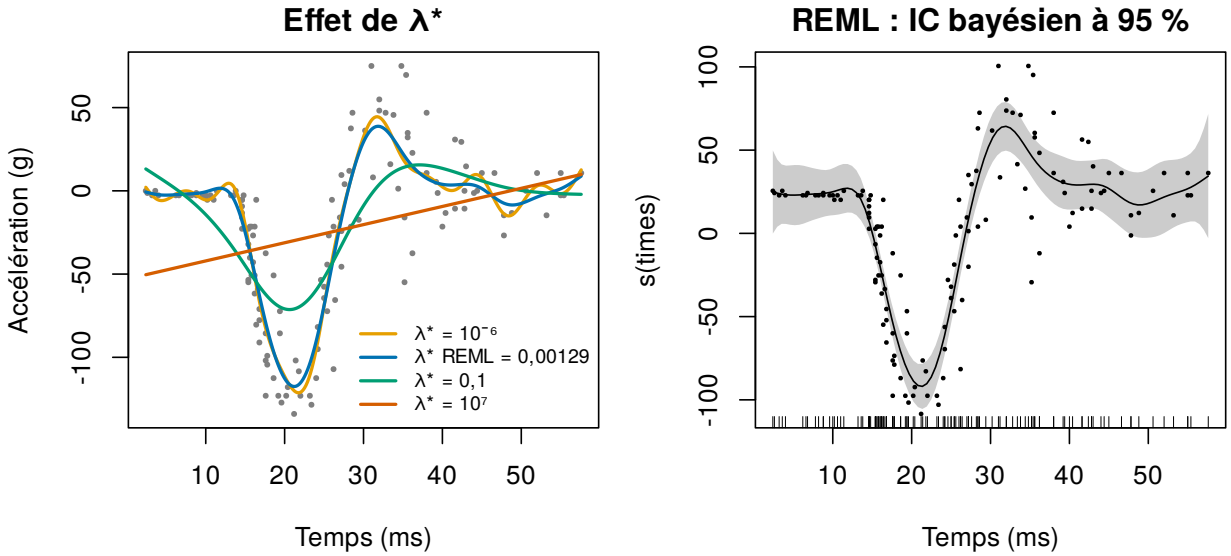


FIGURE 2 – Données `mcycle`. Gauche : ajustements pour quatre valeurs de λ^* (la courbe bleue correspond à la valeur estimée par REML). Droite : estimation REML de la fonction lisse (centrée) avec résidus partiels et intervalle bayésien à 95 %.

Remarque 11.2. La variance de `accel` n'est manifestement pas constante (faible avant 15 ms, élevée ensuite). Les bandes de la figure 2 sont donc trop larges au début et trop étroites ensuite ; un modèle de localisation-échelle (`family = gaulss`) serait plus approprié. Cela n'affecte pas le propos de l'exemple.

Exemple 11.3 (GAMM logistique avec effet aléatoire de sujet)**.** On simule une étude longitudinale de structure semblable à celle du devoir : 300 sujets, visites prévues aux trois mois de 0 à 36 mois (avec un décalage aléatoire de ± 1 mois et un abandon après 5 à 13 visites), un facteur de groupe à trois niveaux fixe dans le temps, une ordonnée à l'origine aléatoire d'écart type 1,2, et une log-cote qui augmente jusqu'à 18 mois puis décroît légèrement : $f(t) = -0{,}8 + 0{,}08\,t - 0{,}11\,(t-18)_+$.

```
set.seed(7360)
n <- 300
temps_vis <- seq(0, 36, by = 3)
f_vrai <- function(t) -0.8 + 0.08 * t - 0.11 * pmax(0, t - 18)
sim <- do.call(rbind, lapply(1:n, function(i) {
  nvis <- sample(5:13, 1)
  t <- temps_vis[1:nvis] + runif(nvis, -1, 1); t[1] <- 0
```

```
  data.frame(id = i, temps = pmax(t, 0),
             groupe = sample(c("A", "B", "C"), 1, prob = c(.5, .3, .2)))
}))
sim$id <- factor(sim$id); sim$groupe <- factor(sim$groupe)
b <- rnorm(n, sd = 1.2)
eta <- f_vrai(sim$temps) + c(A = 0, B = 0.5, C = 0.9)[as.character(sim$groupe)] + b[sim
    $id]
sim$y <- rbinom(nrow(sim), 1, plogis(eta))        # 2720 observations

fit_re <- gam(y ~ groupe + s(temps) + s(id, bs = "re"), family = binomial,
              data = sim, method = "REML")
summary(fit_re)
```

```
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.003942   0.118817  -0.033  0.97353
groupeB      0.469512   0.188366   2.493  0.01268 *
groupeC      0.572698   0.214579   2.669  0.00761 **
Approximate significance of smooth terms:
             edf  Ref.df Chi.sq p-value
s(temps)   3.847   4.745  106.9  <2e-16 ***
s(id)    206.865 297.000  552.7  <2e-16 ***
-REML = 1673.4  Scale est. = 1         n = 2720
```

```
gam.vcomp(fit_re)
```

```
            std.dev       lower      upper
s(temps) 0.02552928 0.009121513 0.07145129
s(id)    1.18417678 1.029968617 1.36147317
```

Quelques constats :

— L'écart type estimé de l'ordonnée à l'origine aléatoire (1,18) est proche de la vraie valeur (1,2). Pour un terme `bs = "re"`, `gam.vcomp` donne directement l'écart type de l'effet aléatoire.

— L'ordonnée à l'origine paramétrique est proche de 0 parce que la contrainte $\sum g(t_{ij}) = 0$ absorbe le niveau moyen de la courbe dans `(Intercept)`.

— L'edf de `s(id)`, environ 207 pour 300 sujets, mesure le degré de rétrécissement des prédictions des b_i ; ce n'est pas un nombre de paramètres au sens habituel.

Comparons ensuite le ML et le REML, puis la spline linéaire par morceaux avec nœud à 18 mois :

```
fit_ml <- update(fit_re, method = "ML")
rbind(REML = fit_re$sp, ML = fit_ml$sp)     # edf de s(temps) : 3,85 (REML), 3,56 (ML)
sim$apres18 <- pmax(0, sim$temps - 18)
fit_pl <- lme4::glmer(y ~ groupe + temps + apres18 + (1 | id), family = binomial, data
    = sim)
summary(fit_pl)$coefficients; VarCorr(fit_pl)
```

```
     s(temps)  s(id)
REML   0.6793 0.7131
ML     0.9402 0.7312
               Estimate  Std. Error   z value     Pr(>|z|)
(Intercept) -0.89080273 0.144461209 -6.166380 6.987100e-10
```

```
groupeB      0.51388783 0.189198062  2.716137 6.604861e-03
groupeC      0.62616643 0.215524542  2.905314 3.668850e-03
temps        0.08642377 0.008468438 10.205397 1.875518e-24
apres18     -0.12063931 0.019168187 -6.293725 3.099361e-10
 Groups Name        Std.Dev.
 id     (Intercept) 1.172
```

Le ML donne un λ plus grand (0,94 contre 0,68), donc une fonction du temps plus lisse (edf de 3,56 contre 3,85), conformément à la tendance annoncée. Le modèle par morceaux estime une pente de 0,086 avant 18 mois et de $0{,}086 - 0{,}121 = -0{,}035$ après, proche des vraies valeurs (0,08 et $-0{,}03$). La figure 3 superpose les deux estimations : la forme apprise par le GAMM suggère bien un changement de pente vers 18 mois, ce qui appuie le choix d'un modèle paramétrique plus simple. Avec `gamm4(y groupe + s(temps), random = (1 | id), family = binomial)`, on obtient des estimations très voisines (groupe B : 0,514 ; groupe C : 0,625 ; écart type de l'effet de sujet : 1,169).

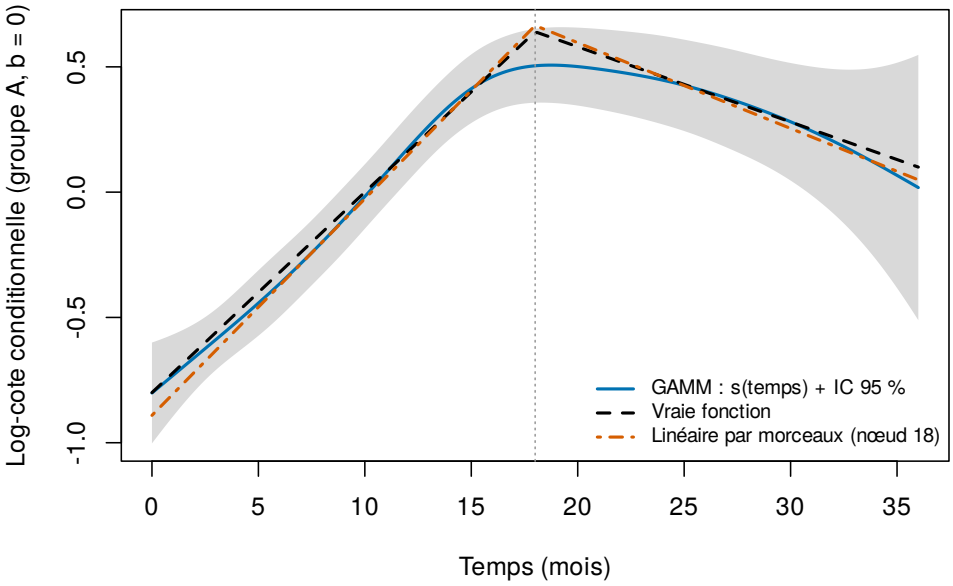


FIGURE 3 – Exemple 11.3 : log-cote conditionnelle de la réponse en fonction du temps pour le groupe A et un sujet d'effet aléatoire nul. Courbe GAMM (REML) avec intervalle bayésien à 95 %, vraie fonction et spline linéaire par morceaux (nœud à 18 mois) ajustée par `glmer`.

Remarque 11.4. Les coefficients de groupe du GAMM et du GLMM sont des effets *conditionnels* (spécifiques au sujet), comme dans tout modèle mixte avec lien non linéaire (séance 3). Ils ne sont pas comparables directement aux coefficients d'un GAM ou d'un GLM marginal ajusté sans effet aléatoire, qui seraient atténués.

12 Synthèse

Grille méthodologique

1. **Objet d'intérêt** : la fonction g (forme de la relation avec le temps ou une covariable continue), et souvent des effets paramétriques $\boldsymbol{\beta}$ ajustés pour cette relation.
2. **Nuisance, sources de biais** : le paramètre de lissage λ ; les effets aléatoires de sujet $\mathbf{b}_i$ et leur covariance $D(\boldsymbol{\eta})$; biais de spécification si l'on impose une forme linéaire à tort, ou biais de rétrécissement de l'estimateur pénalisé.
3. **Approche** : base de splines de grande dimension plus pénalité de rugosité ; représentation du modèle pénalisé comme GLMM, de sorte que λ devient une composante de variance.
4. **Critère** : vraisemblance pénalisée (3) pour $\boldsymbol{\gamma}$; GCV/UBRE ou approximation de Laplace de la vraisemblance (PL) ou de la vraisemblance restreinte (PL_r) pour λ et $\boldsymbol{\eta}$.
5. **Calcul** : P-IRLS (boucle interne) imbriqué dans une méthode de Newton sur $\log\lambda$ (boucle externe) ; ou méthodes de `lme4` via `gamm4`.
6. **Inférence** : loi a posteriori normale approximative de $\boldsymbol{\gamma}$ (intervalles bayésiens, couverture « à travers la fonction ») ; tests de Wald approximatifs ; tests du rapport de vraisemblance par ML pour comparer des parties fixes.

À retenir de la séance

— $g(t) = \boldsymbol{\gamma}^\top \boldsymbol{\psi}(t)$ et $\int g''^2 = \boldsymbol{\gamma}^\top S \boldsymbol{\gamma}$: un GAM est un GLM pénalisé. La base doit être assez grande ; la pénalité règle la régularité.

— Pour λ fixé, P-IRLS = moindres carrés pondérés pénalisés itérés sur la pseudo-réponse.

— La GCV choisit λ par prédiction ; elle suppose l'indépendance et peut sous-lisser.

— Le GAM est un GLMM (effets aléatoires de variance $1/\lambda$) ; ML et REML estiment λ comme une composante de variance, et s'étendent naturellement aux GAMM.

— ML $\Rightarrow$ variance sous-estimée $\Rightarrow$ λ plus grand $\Rightarrow$ courbe plus lisse.

— L'edf mesure la complexité effective ; les intervalles de mgcv sont bayésiens ; les tests sont approximatifs.

— `gam` + `bs = "re"` : ordonnées à l'origine aléatoires faciles, pentes aléatoires indépendantes seulement ; pour des pentes corrélées, `gamm4` ou une spline linéaire par morceaux dans `glmer`.

Exercices

Exercice 4.1 (Matrice de pénalité ; inspiré de GAM_2026, H26)**.** Soit la base de puissances tronquées de degré 3 à un nœud $\kappa \in (0,1)$ sur $[0,1]$: $\boldsymbol{\psi}(t) = \{1, t, t^2, t^3, (t-\kappa)_+^3\}^\top$.

(a) Calculez explicitement la matrice S de la proposition 3.1.

(b) Vérifiez que son rang est 3 et identifiez son noyau.

(c) Avec R, comparez au résultat de `smoothCon(s(t, bs = "cr", k = 5), ...)$S` et expliquez

pourquoi les deux matrices diffèrent tout en ayant le même rang.

Exercice 4.2 (Validation croisée)**.** Montrez que, pour un lisseur linéaire, $V_g(\lambda) \approx V_o(\lambda)$ lorsque les éléments diagonaux A_{ii} sont tous voisins de leur moyenne, et que V_g est invariant si l'on remplace $(\mathbf{y},\Psi)$ par $(Q\mathbf{y},Q\Psi)$ pour une matrice orthogonale Q, ce qui n'est pas le cas de V_o.

Exercice 4.3 (GAM gaussien et équations de Henderson)**.** Dans le modèle $\mathbf{y} = X_F\boldsymbol{\beta}_F + Z\mathbf{u} + \boldsymbol{\varepsilon}$, $\mathbf{u} \sim \mathcal{N}(\mathbf{0},\sigma_u^2 I)$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\sigma^2 I)$:

(a) écrivez les équations du modèle mixte de Henderson et montrez qu'elles coïncident avec les équations normales des moindres carrés pénalisés avec $\lambda^* = \sigma^2/\sigma_u^2$;

(b) déduisez-en que le BLUP de $Z\mathbf{u}$ est la partie pénalisée de la spline estimée ;

(c) montrez que $\text{edf} = M_0 + \sum_j d_j/(d_j + \lambda^*)$ dans ce cas.

Exercice 4.4 (Deux fonctions lisses ; d'après les notes H26 de la séance 3)**.** Wu (p. 76, éq. 2.64–2.65) traite un modèle additif avec deux fonctions lisses. Écrivez la log-vraisemblance pénalisée de ce modèle avec deux paramètres de lissage λ_1,λ_2, puis sa représentation GLMM avec deux vecteurs d'effets aléatoires indépendants de variances $1/\lambda_1$ et $1/\lambda_2$. Écrivez l'approximation de Laplace de la vraisemblance restreinte $PL_r(\lambda_1,\lambda_2)$.

Exercice 4.5 (Étude de simulation : ML, REML et GCV)**.** En vous inspirant de la séance 2, planifiez et réalisez une petite étude de simulation qui compare les edf et l'erreur quadratique intégrée $\int\{\hat{g}(t) - g(t)\}^2\,\mathrm{d}t$ des méthodes `GCV.Cp`, `REML` et `ML` de `gam`, pour $g(t) = \sin(2\pi t)$, $n \in \{30,100,400\}$ et deux niveaux de bruit. Ajoutez ensuite au modèle un facteur à 12 niveaux (effets fixes) : l'écart entre ML et REML change-t-il ? Interprétez à la lumière de la proposition 7.2.

Exercice 4.6 (GAMM et choix du nœud ; inspiré du devoir 1, H26)**.** Avec les données simulées de l'exemple 11.3 (ou en les régénérant avec une autre graine) :

(a) ajustez le GAMM avec `s(temps, by = groupe)` et testez approximativement si la forme de l'évolution temporelle diffère entre les groupes (ajustements par ML) ;

(b) pour des nœuds $\kappa \in \{6,9,\ldots,30\}$, ajustez la spline linéaire par morceaux avec `glmer` et tracez la log-vraisemblance profilée en fonction de κ ;

(c) expliquez pourquoi la valeur-p du test de $\beta_2 = 0$ au nœud qui maximise la vraisemblance n'a pas son interprétation habituelle, et proposez une façon (p. ex. par simulation sous H_0) d'obtenir un test valide.

Exercice 4.7 (Niveau doctorat)**.** Complétez l'exercice guidé de l'encadré « Laplace restreinte et quasi-vraisemblance pénalisée ». Puis, par simulation (réponse binaire, 5 observations par sujet, écart type de l'effet aléatoire égal à 2), comparez l'estimation de l'écart type de l'effet de sujet par `gamm` (PQL), par `gam(..., s(id, bs = "re"), method = "REML")` et par `gamm4`.

Références

— Breslow, N. E. et Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.

— Craven, P. et Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 377–403.

— de Boor, C. (1978). *A Practical Guide to Splines*. Springer.

— Eilers, P. H. C. et Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121.

— Green, P. J. et Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.

— Lin, X. et Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society B*, 61, 381–400.

— Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, 83, 1134–1143.

— Ruppert, D., Wand, M. P. et Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

— Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society B*, 73, 3–36.

— Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100, 221–228.

— Wood, S. N. (2017). *Generalized Additive Models : An Introduction with R* (2$^{\text{e}}$ éd.). Chapman & Hall/CRC.

— Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC, § 2.5–2.6.

— Bureau, A. (2026). *Modèles additifs généralisés (mixtes)*. Document complémentaire, STT-7360, Université Laval.