Modèle de régression sur données longitudinales: Application à des données médicales à faible résolution

Christophe BOTELLA

1er septembre 2016

Table des matières

Résumé

Présentation des données et leurs spécificités. Formalisation du problème statistique. Évaluation d'une approche existante pour la modélisation de données longitudinales à domaine variable et des temps de survie. Proposition d'un modèle. Analyse de performance comparée par simulation.

1 Introduction

Les modèles dits de "survie" mettant en relation une durée jusqu'à l'apparition d'un événement avec des variables vitales de patients, comme le fameux modèle de Cox [Cox, 1992], ont été très explorés dans les dernières décennies. Le stockage de mesures vitales évoluant au cours du temps, comme la glycémie ou la tension artérielle, dans les enregistrements électroniques de santé (EHR) et entrepôts de données hospitaliers est un phénomène plus récent. Ces bases de données fournissent des données extrêmement utiles pour la recherche médicale et encore sous exploitées. Accessibles et incluses dans des modèles de prédiction, elles sont susceptibles d'améliorer la prise en charge des patients. Cependant, la quantité et la diversité croissante d'informations regroupées dans les dossiers électroniques posent des problèmes d'exploitation pratique pour les praticiens. Les professionnels de la santé manifestent un intérêt croissant pour les approches prédictives à partir de ces données, dans le but d'appuyer les décisions médicales. Dans ce contexte, les modèles de régression qui établissent un lien entre les données vitales longitudinales et l'apparition de complications médicales sont d'un grand intérêt et seront au centre de notre étude.

On se place ici dans le cadre de données dîtes de ré-hospitalisation que nous décrivons ci-après. On s'intéresse à une cohorte de N patients. Chaque patient $i \in \{1...n\}$ a été hospitalisé pendant une durée notée T_i suite à un événement médical lié au problème étudié (une opération, un soin etc...). Pendant cette durée T_i, on mesure D signes vitaux fluctuants (tension, glycémie, etc.) à des instants discrets. Ces instants de mesure ne sont pas forcément les même entre les patients et entre les signes vitaux d'un patient. Pour un patient i et un signe vital j, on note t_j^i le vecteur des instants de mesures et X_j^i les valeurs mesurées. Parallèlement, des données supposées constantes sur l'échelle de temps de T_i (âge, poids, taille etc.) sont relevées et notées $Z_i \in \mathbb{R}^p$. Le patient sort de l'hôpital, mais il est parfois ré-admis précocement pour une raison qui peut être liée à une prise en charge inadéquate durant son séjour. Dans ce cas, la donnée indiquant la durée entre sa sortie et la ré-hospitalisation est notée τ_i. Comme la ré-admission précoce n'est pas systématique (ou pas connue), on considère une censure à droite des données en indiquant par $\delta_i = 1$ le fait qu'on aie observé une ré-admission pour le patient i, ou $\delta_i = 0$ autrement. Notons C_i le temps de censure du patient i. C_i est la durée entre la dernière sortie de l'hôpital du patient i et le moment où il quitte la cohorte (on ne peut alors plus avoir aucune information sur i). On note alors le temps censuré $\tau_i^c = min(\tau_i, C_i)$. Une représentation des données est fournie par la figure 1 : Ici un signe vital est mesuré pendant les séjours respectifs des patients 1 et 2. On représente alors la durée jusqu'à l'événement/censure de chacun avec $\tau_1^c = \tau_1$ et $\tau_2^c = C_2$.

Dans de nombreux cas, une prise en charge mieux adaptée au patient (durée du séjour, doses de médicaments) pourrait diminuer le taux de ré-hospitalisation. Par exemple, une analyse de données d'hospitalisation pour Crises Vaso-Occlusives (CV0) liées à la Drépanocytose aux Etats-Unis montre un taux de ré-hospitalisation précoce (proportion des patients ré-admis dans l'hôpital dans les deux semaines suivant le séjour) de 22% sur plus de 10 000 séjours, avec une fréquence plus haute chez les patients jeunes Brousseau *et al.* [2010]. Une étude plus générale du MedPac [Medicare-Payment-Advisory, 2007] révèle que 17.6% des hospitalisations aux Etats-Unis donneraient lieu à des ré-admissions en moins de 30 jours, et que 76% de celles-ci seraient potentiellement évitables. Pour parvenir à mieux adapter la prise en charge des patients, l'analyse et la prédiction à partir des données longitudinales est cruciale. Elles caractérisent l'évolution de l'état de santé du patient suite à une intervention et peuvent alerter précocement des dangers de ré-admission. Prenons le cas simple de la douleur qui est souvent mesurée au cours des séjours hospitaliers. La valeur mesurée est très subjective, mais la tendance donne une information très importante quant au danger de ré-hospitalisation précoce qui est reconnue parmi les médecins.

Pour clarifier le problème, établissons les notations pour les données que nous avons en main :

$$\mathcal{D} = \left\{ t^i := (t_j^i)_{j \in [|1,D|]}, X^i := (X_j^i(t_j^i))_{j \in [|1,D|]}, Z_i \in \mathbb{R}^p, T_i \in \mathbb{R}, \tau_i^c \in \mathbb{R}, \delta_i \in \{0,1\}, i \in [|1,N|] \right\}$$

où, pour tout patient $i \in [|1, N|]$:

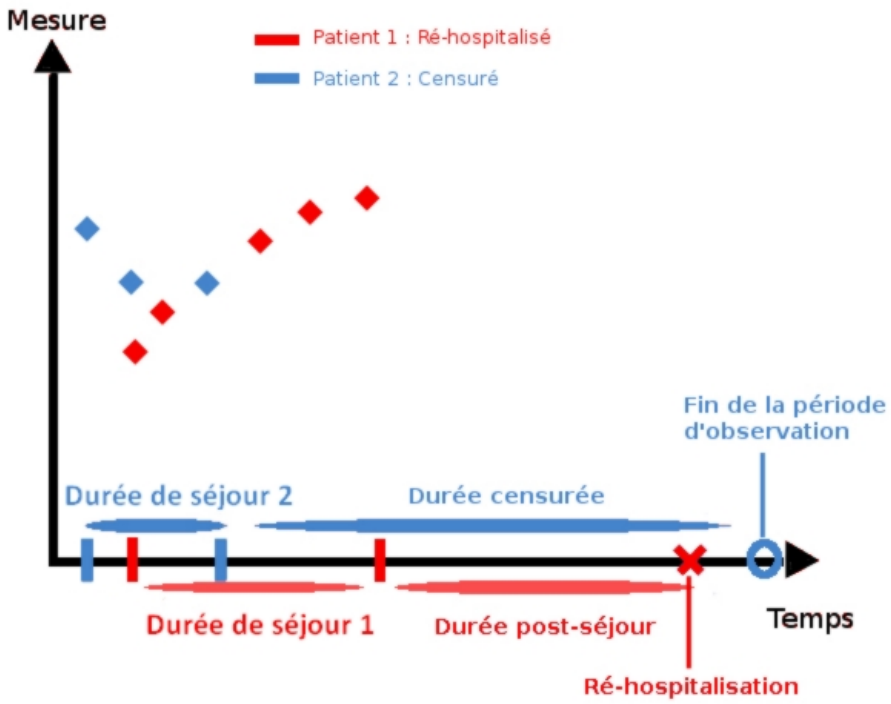

FIGURE 1 – Représentation de données de ré-hospitalisation temporelles censurées

— T_i est la durée de mesure (durée de prise en charge hospitalière).

— $\forall j \in [|1, D|], t_j^i = (t_{j1}^i, ..., t_{jm_{ij}}^i)^T \in \mathbb{R}^{m_{ij}}$ est le vecteur des instants d'observations du signe de vie j du patient i dans l'intervalle $[0, T_i]$ et m_{ij} le nombre d'instants. On utilise la notation t^i pour la liste des vecteurs $t_j^i, j \in [|1, D|]$.

— $\forall j \in [|1, D|], X_j^j(t_j^i) = (X_j^i(t_{j1}^i), ..., X_j^i(t_{jm_{ij}}^i))^T \in \mathbb{R}^{m_{ij}}$ est le vecteur des valeurs associées. On utilise la notation X^i pour la liste des vecteurs $X_j^i(t_j^i), j \in [|1, D|]$.

— Z_i est le vecteur des variables scalaires du patient.

— $\tau_i^c = min(\tau_i, C_i)$ est le temps censuré d'apparition d'un événement (ré-admission hospitalière).

— δ_i est l'indicateur de non-censure, $\delta_i = 1$ si $\tau_i^c = \tau_i$ ou O sinon.

Plus généralement dans tout ce qui suit, et sauf indication contraire, le numéro du patient sera indexé par i et le signe vital par j. Une notations qui n'arbore que l'indice du patient fait référence à la liste de l'ensemble des signes vitaux du patient. De même, si elle n'arbore aucun indice, elle renvoie à la liste de l'ensemble des patients et signes vitaux.

On souhaiterait analyser le lien entre les temps de ré-hospitalisation censurés τ_i^c, δ_i et les variables Z_i et X^i grâce à un modèle de survie qui valorise l'information des variables temporelles X^i afin de prédire le risque de ré-hospitalisation précoce des patients. De nombreux exemples de recherche appliquée existent autour des problèmes de ré-hospitalisation. Baillie *et al.* [2013] présente l'implémentation d'un outil logiciel utilisant un modèle d'apprentissage sur une cohorte rétrospective pour prédire et cibler des patients dont le risque de ré-hospitalisation avant 30 jours est élevé. L'expérimentation sur une cohorte prospective a montré une stabilité de la capacité prédictive du modèle par rapport aux jeu d'entraînement, même si elle n'a pas permis de réduire le taux de ré-hospitalisation. Une comparaison de modèles de prédictions linéaires utilisés en pratique à d'autres modèles statistiques non-linéaires comme les forêts aléatoires ou les réseaux neuronaux profonds est menée dans Futoma *et al.* [2015] sur données réelles. Amarasingham *et al.* [2010] s'intéresse à des données de ré-hospitalisation pour

cause d'arrêt cardiaque et propose un modèle basé sur des variables cliniques et non-cliniques avec de très bons résultats de prédiction pour la ré-admission et le décès dans les 30 jours suivant la sortie du patient de l'hôpital. Ces résultats nous confortent dans l'espoir que le risque de ré-admission puisse être estimé de manière précise dans beaucoup de cas médicaux en s'intéressant aux bons facteurs de risques et avec des modèles adéquats. Cependant, tous les modèles précédemment cités se restreignent ou se ramènent à des variables explicatives de type scalaires, que nous avons notées Z_i.

L'inclusion de variables évolutives comme les X_j^i dans des modèles d'analyse de survie est un sujet plus récent et théorique (Wu *et al.* [2011]). Ces modèles intègrent des variables fonctionnelles et s'inscrivent dans la lignée des travaux sur la régression fonctionnelle (Ramsay [2006], Ferraty & Vieu [2006]). L'idée est d'utiliser l'information contenue dans la trajectoire des variables fonctionnelles X^i (signes de vie) mesurées sur un intervalle réel fini (le temps). Or, la majorité de la trajectoire est cachée puisque, techniquement, nous ne l'observons que par des mesures à des instants discrets, éventuellement bruitées. Le modèle de régression fonctionnelle linéaire répond à ces problématiques de manière paramétrique (Ramsay [2006],Goldsmith *et al.* [2012]), mais d'autres modèles non-paramétriques ont été proposés (Ferraty & Vieu [2006]) par des méthodes à noyaux. Cependant, ils supposent un intervalle fixe de mesure pour les variables fonctionnelles. Dans le cas présent, où la durée de suivi de chaque patient est variable, des développements ont été amenés par Gellar *et al.* [2014] avec le modèle de régression fonctionnelle à domaine variable (VDFR) qui s'appuie sur le cadre de la régression fonctionnelle pénalisée . Ces modèles s'intéressent à des variables fonctionnelles dont la mesure doit respecter certaines hypothèses, dont une haute densité d'échantillonnage, notion que nous définirons plus loin. Cependant, ce dernier modèle fournit une base intéressante pour répondre à notre problématique. De telles hypothèses sont en contraste avec notre situation : Dans le cas de signes vitaux mesurés lors de consultations, ou de séjours hospitaliers, la densité d'échantillonnage est faible par rapport au potentiel de fluctuation. Nous étendons le modèle VDFR à un modèle joint de régression fonctionnelle semi-paramétrique basé sur des processus gaussiens. Les processus gaussiens sont un outil probabiliste intéressant pour modéliser des variables fonctionnelles aléatoires. Leurs propriétés et les méthodes d'inférence sont revues dans Rasmussen [2006]. Nous commençons par présenter les spécificités des données de ré-hospitalisation, nous testons la performance du modèle VDFR dans le cadre d'une densité d'échantillonnage faible et d'erreur sur les variables fonctionnelles, puis, nous proposons d'étendre ce modèle avec le schéma d'inférence fonctionnelle jointe par des processus gaussiens. Enfin, nous mettons à l'épreuve ce modèle avec différentes méthodes d'inférence sur des données simulées et réelles.

2 Particularités des données longitudinales

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. Une définition assez générale d'une variable aléatoire fonctionnelle $\mathcal{F} : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^T$ introduite dans Ferraty & Vieu [2006] est une variable aléatoire à valeur dans un espace de dimension infinie. On peut donc identifier le processus $\mathcal{F}$ à l'ensemble des variables aléatoires locales : $\{\mathcal{F}(t), t \in \mathcal{T}\}$. Nous nous intéressons ici à des **variables fonctionnelles temporelles (VFT)**, dont l'ensemble de définition est un intervalle réel que l'on prend de la forme $[0, T] \subset \mathbb{R}$ où T est la durée de suivi. Une réalisation F de la VFT $\mathcal{F}$ est une donc une fonction de $[0, T]$ sur $\mathbb{R}$. En considérant un patient i et un signe vital j, nous observons un ensemble fini de mesures à des temps discrets de la fonction F_j^i qui incarne le signe vital de ce patient sur sa période de suivi $[0, T_i]$. Les mesures collectées sur l'ensemble de la cohorte constitue notre **jeu de données fonctionnel observé**.

On peut à présent introduire quelques spécificités liées à nos données fonctionnelles observées :

— **La densité d'échantillonnage ou résolution** d'un jeu de données fonctionnel observé : C'est le nombre moyen de mesures réalisées par unité de temps sur l'intervalle de définition de la variable fonctionnelle. Si l'on a aucune idée des fluctuations de la VFT, cet indicateur ne nous apprend rien d'utile. En revanche, à la connaissance des ordres de grandeurs des variations de la fonction mesurée, on est en mesure de juger si la densité d'échantillonnage est grande ou

faible. C'est un facteur clé de la performance de la régression fonctionnelle [Ramsay, 2006]. Lorsque la résolution est faible, la méthode de traitement qui infère une fonction continue à partir de la VFT observée doit être adaptée aux signal mesuré et combler les zones d'ombre de l'échantillonnage sans induire de biais.

— **La variabilité des domaines de mesure** : La variabilité des durées de séjours hospitaliers entraîne une variabilité de l'intervalle de définition des VFTs dans les jeux de données fonctionnels auxquels on s'intéresse. Pourtant, les méthodes qui permettent de prendre en compte cette spécificité sont rares.

— **L'erreur sur la variable** : Il s'agit d'une erreur commise sur la mesure par rapport à la valeur réelle du signe vital observé.

On comprend que la densité d'échantillonnage et les erreurs de mesures rendent impossible l'accès direct à la VFT. Lorsqu'on souhaite utiliser les VFT comme variables dans un modèle de survie pour prédire les temps jusqu'à l'événement, on doit intégrer une procédure d'inférence des VFT, de manière préalable ou simultanée à l'inférence de la relation entre VFT et données de survie.

Remarque : Dans la suite on confondra, sauf remarque contraire, la notation d'une variable aléatoire (réelle ou fonctionnelle) et de son observation.

3 Proposition d'un modèle joint adapté à une faible densité d'échantillonnage fonctionnel

Rappelons que dans le modèle original de Cox, la loi du temps de survie est de forme inconnue et on ne cherche pas à l'estimer, mais plutôt à pouvoir comparer les risques relatifs des patient, les uns par rapport aux autres. Nous voulons être en mesure de prédire les chances de survie à 5 ans, la durée de survie ainsi la fiabilité de cette prédiction, à la connaissance de variables vitales d'un patient. Cela suppose de modéliser et d'estimer la loi du temps de survie. **Nous faisons l'hypothèse que nous connaissons une transformation du temps de survie dont on sait que la loi est gaussienne**.

Dans un premier temps, nous éluderons volontairement la censure des données. On prend un individu générique avec une VFT pour simplifier l'écriture sans perte de généralité.

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. On note $F : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^{\mathcal{T}}$ une fonction aléatoire générique. Soit $m \in \mathbb{N}$, $t = (t_1, ..., t_m)^T \in \mathcal{T}^m$ le nombre de mesures, $\zeta \in \mathbb{R}^m$ un vecteur aléatoire. Soit $\gamma, \sigma \in \mathbb{R}$, $k : \mathbb{R}^2 \to \mathbb{R}^+$ une fonction continue, positive et symétrique dépendant de paramètres $\theta \in \Theta$.

Modélisation d'une VFT par un processus gaussien. On considère suivant pour la VFT :

$$X(t) = F(t) + \zeta \quad \textbf{(1)}$$

où $X(t) = (X(t_1), ..., X(t_m))^T$ et $F(t) = (F(t_1), ..., F(t_m))^T$. Si on considère une réalisation de $X(t)$ et qu'on souhaite estimer la réalisation du processus F qui se cache derrière le vecteur $X(t)$, nous pouvons utiliser le schéma de régression sur processus gaussien détaillé dans Rasmussen [2006], dont le principe est décrit. Pour ce modèle, on considère les hypothèses suivantes :

Hypothèses 1.
1. $F \sim \mathcal{GP}(0_{\mathbb{R}^{\mathcal{T}}}, k)$ *est un processus gaussien de moyenne constante et nulle, et de fonction de covariance* k.
2. $\zeta \sim \mathcal{N}(0_m, \gamma^2 I_m)$.

Remarque : On prend ici un processus gaussien de moyenne nulle pour simplifier les notations mais les propriétés qui suivront se généralisent très simplement à n'importe quelle fonction moyenne,

en affectant simplement l'espérance des lois décrites. En particulier, on s'intéressera par la suite à des processus de moyenne constante.

Pour faire le parallèle avec les données, $X(t)$ représente, comme définit dans l'introduction, les mesures du signe vital de l'individu aux instants $t_1, ..., t_m$. F représente le signe vital réel, c'est à dire non perturbé par une erreur de mesure. Même en supposant que

On suppose que cette fonction est une trajectoire générée par un processus gaussien. Cette approche est qualifiée de semi-paramétrique, car les seuls paramètres du modèle sont ceux de la fonction de covariance k du processus gaussien, et γ qui quantifie l'erreur de mesure. L'avantage de cette modélisation semi-paramétrique est qu'elle utilise peu de paramètres pour décrire un grand ensemble fonctions. Ceci est possible car on s'appuie sur les points observés et sur une hypothèse quant à la forme de la fonction de covariance, qui régit la régularité de la courbe. Résumer ainsi la fonction avec peu de paramètres est intéressant ici, car, dans un contexte de faible densité d'échantillonnage, une modélisation paramétrique pourrait entraîner un sur-ajustement aux données. Si des méthodes de pénalisations existent [Goldsmith *et al.* , 2012] pour corriger ce problème, le modèle de régression sur processus gaussien a un autre avantage qui le rend plus intéressant : le choix de la fonction de covariance nous permet d'intégrer de la connaissance à priori sur la courbe modélisée. Par exemple, dans le cas de données médicales, le rythme circadien affecte grandement la plupart des signes vitaux des patients. Nous pouvons intégrer cette connaissance sur le signal en ajoutant un noyau de covariance périodique, et on aura alors de bonnes chances de capter la périodicité du signal, même avec peu de points d'observation. En ce sens, la modélisation **(1)** semble bien répondre à la faible densité échantillonnage, et permet une flexibilité de modélisation du signal basée sur la connaissance du phénomène. L'objectif de ce travail de stage va consister à intégrer cette modélisation des VFT au cadre de la régression fonctionnelle linéaire pénalisée (Ramsay [2006],Goldsmith *et al.* [2012]) en conjecturant qu'elle peut amener de meilleurs résultats que les modélisations paramétriques, comme celle qui est proposée dans Gellar *et al.* [2014].

Modélisation du temps de survie. Soit τ une variable aléatoire réelle correspondant à la durée entre la sortie de l'hôpital de notre individu et sa ré-hospitalisation. Nous ne connaissons pas la loi de τ mais nous faisons l'hypothèse que nous connaissons une transformation de τ dont la loi est gaussienne. Différentes formes de transformations ont été étudiées dans la littérature dans le but de satisfaire au mieux cette hypothèse, comme présenté dans Barrett & Coolen [2013]. Soit $Z, \alpha \in \mathbb{R}^p$, soit $T \in \mathbb{R}^+$ tel que $\mathcal{T} = [0, T]$. On intègre alors la variable fonctionnelle F dans le modèle de régression linéaire fonctionnelle sur le label Y :

$$Y = Z^T\alpha + \frac{1}{T}\int_0^T F(s)\beta(s,T)ds + \epsilon \quad \textbf{(2)}$$

Malheureusement, même en supposant que θ soit connu et qu'il n'y aie pas d'erreur de mesure sur F, nous ne pouvons pas inférer la valeur exacte $F(s)$ pour $s \notin \{t_1, ..., t_m\}$ selon le modèle **(1)**. Par conséquent, quand bien même nous aurions en mains les bons paramètres des modèles **(1) et (2)**, et conditionnellement à des observations $(X(t_1), ..., X(t_m))$ nous serions incapables de prédire exactement Y, à cause des intervalles inobservés de F. Y est d'après le modèle, une variable aléatoire qui dépend de la variable cachée F. Aussi, nous aimerions au moins connaître son espérance et sa variance, elles nous seraient utiles d'un point de vue prédictif. Nous allons voir que l'a priori processus gaussien sur F fait que l'on connaît sa loi, de laquelle on va déterminer conjointement celle de $(X(t_1), ..., X(t_m))$ et Y, qui sont liées. Le modèle **(2)** inclue les hypothèses **1** et **2** ci-dessous.

Hypothèses 2.
1. *Il existe $g : \mathbb{R}^+ \to \mathbb{R}$ une application strictement monotone connue telle que $g(\tau)$ suit une loi normale. On note ainsi la transformation de temps de survie $Y := g(\tau)$, cette variable incarnera donc le label de notre modèle de survie.*
2. *$\beta \in \mathbb{R}^{\mathbb{R}^2}$ une fonction continue et bornée sur la surface $\{(s,t) \,/\, t \in \mathcal{T}, inf(\mathcal{T}) \leq s \leq t\}$.*

3. $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Il en découle la propriété suivante sur le loi jointe de (Y, X)

Propriété 1. *Avec les définitions précédentes et sous les hypothèses 1. et 2. on a :*

$$\begin{pmatrix} \frac{1}{T}\int_0^T F(s)\beta(s,T)ds \\ X(t_1) \\ \vdots \\ X(t_m) \end{pmatrix} \sim \mathcal{N}\left(0_{m+1}, \begin{pmatrix} \frac{1}{T^2}\int_{[0,T]^2} k(s,t)\beta(s,T)\beta(t,T)dsdt & \frac{1}{T}\int_{[0,T]} K_*^T(s)\beta(s,T)ds \\ \frac{1}{T}\int_{[0,T]} \beta(s,T)K_*(s)ds & K_t + \gamma^2 I_m \end{pmatrix}\right)$$

$$Où\ \forall s \in [0,T], K_*(s) = (k(s,t_1), ..., k(s,t_m))^T\ et\ K_t = \begin{pmatrix} k(t_1,t_1) & & \\ \vdots & \ddots & \\ k(t_m,t_1) & \cdots & k(t_m,t_m) \end{pmatrix} \in \mathcal{M}_{mm}(\mathbb{R}).$$

La preuve est donnée dans l'appendice.

Alors, la loi jointe de $(Y, X(t_1), ..., X(t_m))^T$ découle directement de la propriété car ϵ suit une loi normale indépendante de $\frac{1}{T}\int_0^T F(s)\beta(s,T)ds$ et de $(X(t_i))_{i\in[|1,m|]}$. Il s'agit de la loi

$$\mathcal{N}\left(\begin{pmatrix} Z^T\alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{T^2}\int_{[0,T]^2} k(s,t)\beta(s,T)\beta(t,T)dsdt + \sigma^2 & \frac{1}{T}\int_{[0,T]} K_*^T(s)\beta(s,T)ds \\ \frac{1}{T}\int_{[0,T]} \beta(s,T)K_*(s)ds & K_t + \gamma^2 I_m \end{pmatrix}\right)$$

Par ailleurs, on obtient le corollaire suivant par la propriété de conditionnement gaussien.

Corollaire 1. *Sous les hypothèses 1. et 2., conditionnellement à $X(t)$, la variable aléatoire réelle Y suit une loi normale avec :*

$$\begin{aligned} \mathbb{E}(Y|X(t)) &= Z^T\alpha + \frac{1}{T}\int_0^T \beta(s,T)\,\mathbb{E}(F(s)|X(t))\,ds \\ &= Z^T\alpha + \frac{1}{T}\int_0^T \beta(s,T)\,K_*(s)^T(K_t+\gamma^2 I_m)^{-1}X(t)ds \\ \mathbb{V}(Y|X(t)) &= \sigma^2 + \frac{1}{T^2}\int_{[0,T]^2} \mathbb{C}ov\left(F(s),F(t)|X(t)\right)\beta(s,T)\beta(t,T)dsdt \\ &= \sigma^2 + \frac{1}{T^2}\int_{[0,T]^2} \left(k(s,t) - K_*(s)^T.(K_t+\gamma^2 I_m)^{-1}.K_*(t)\right)\beta(s,T)\beta(t,T)dsdt \end{aligned}$$

Remarque, notation : Les espérances, variances et covariances conditionnelles qui sont écrites ici dépendent implicitement des paramètres $\mathcal{P}$ mais nous ne le faisons pas apparaître dans les écritures ici et dans ce qui suit. De plus, $X(t)$, le vecteur aléatoire des observations du signe vital aux temps de mesure $t = (t_1, ..., t_m)$, sera souvent noté X lorsqu'un seul signe vital est considéré. De la même manière, et sans provoquer d'ambiguïté selon les contextes de notations, X représentera aussi l'ensemble des signes vitaux observés, X^i les signes vitaux du patient i, X_j^i le j^{eme} signe vital du patient i, à leurs temps d'observations respectifs.

Remarque, Consistance de $s \to \mathbb{E}(F(s)|X(t))$: Dans le schéma de régression sur processus gaussien du modèle **(1)**, qui est décrit par Rasmussen [2006], des résultats de consistance de l'estimateur $s \to \mathbb{E}(F(s)|X(t))$ ont été apportés [Choi & Schervish, 2007] lorsque le nombre d'observations

$m \to +\infty$ et que les temps d'observations $t_1, ..., t_m$ sont échantillonnés uniformément sur un intervalle fini.

4 Cas de plusieurs variables fonctionnelles indépendantes

A présent, on étend le modèle **(2)** au cas de $D \in \mathbb{N}$ variables fonctionnelles $X^1, ..., X^D$ et on considère N observations correspondant à N patients pour lesquels on a mesuré ces variables fonctionnelles, des variables scalaires et un temps de survie.

Hypothèses 3. *1. Les N observations sont indépendantes.*

2. Les D variables fonctionnelles sont indépendantes.

L'hypothèse **2.2**, bien que courante, peut être non-valable pour des données médicales, où les signaux vitaux ont des variations et des cycles souvent corrélés. Des développements pourraient être amenés à l'avenir en assumant une dépendance des variables fonctionnelles. Rappelons les notations des données afin d'introduire le modèle :

$$\mathcal{D} = \left\{ t^i := (t^i_j)_{j \in [|1,D|]}, X^i := (X^i_j(t^i_j))_{j \in [|1,D|]}, Z_i \in \mathbb{R}^p, T_i \in \mathbb{R}, Y_i \in \mathbb{R}, i \in [|1, N|] \right\}$$

où pour tout patient $i \in [|1, N|]$:

— $T_i \in \mathbb{R}^+$ est la durée de mesure (durée de suivi hospitalier) associée à l'intervalle de mesure $\mathcal{T}_i = [0, Ti]$.

— $\forall j \in [|1, D|], t^i_j = (t^i_{j1}, ..., t^i_{jm_{ij}})^T \in \mathbb{R}^{m_{ij}}$ est le vecteur des temps d'observations de la variable fonctionnelle j (signe de vie) du patient i dans l'intervalle $[0, T_i]$. On utilise la notation t^i pour la liste des vecteurs $t^i_j, j \in [|1, D|]$.

— $\forall j \in [|1, D|], X^j_j(t^i_j) = (X^i_j(t^i_{j1}), ..., X^i_j(t^i_{jm_{ij}}))^T \in \mathbb{R}^{m_{ij}}$ est le vecteur des valeurs associées. On utilise la notation X^i pour la liste des vecteurs $X^i_j(t^i_j), j \in [|1, D|]$.

— Z_i est le vecteur des variables scalaires du patient.

— $Y_i = g(\tau_i)$ où τ_i est le temps d'apparition de l'événement (ré-hospitalisation) du patient i.

Remarque : Notons qu'on a modifié ici l'hypothèse **1.1** du modèle **(1)** en supposant cette fois-ci que les processus gaussiens F^i_j sont de moyenne constante. Cela retranscrit le fait que les signaux vitaux sont de façon évidente non-nuls en moyenne si on les observe au-delà de l'intervalle de mesure. Ainsi, il est plus correct d'ajouter la moyenne du processus comme un paramètre supplémentaire plutôt que de l'estimer à priori avec la moyenne des valeurs observées. Ceci introduirait un biais d'estimation au vu de la modélisation choisie.

On s'intéresse au modèle suivant, $\forall i \in [|1, N|]$:

$$\textbf{(3)} \begin{cases} Y_i = Z_i^T \alpha + \dfrac{1}{T_i} \displaystyle\sum_{j=1}^{D} \int_0^{T_i} F^i_j(s) \beta_j(s, T_i) ds + \epsilon_i \\ X^i_j(t^i_j) = F^i_j(t^i_j) + \zeta^i_j \ , \ \forall j \in [|1, D|] \end{cases}$$

Où :

— $\alpha \in \mathbb{R}^p$ est le vecteur des paramètres des variables scalaires

— $\beta_j : \mathbb{R}^{+2} \to \mathbb{R}$ est la fonction paramètre associée à la j^{eme} variable fonctionnelle.

— $\epsilon_i \in \mathbb{R}$ et $\zeta_j^i \in \mathbb{R}^{m_{ij}}$ sont des erreurs iid, gaussiennes centrées de variances respectives $\sigma^2 \in \mathbb{R}^+, \gamma_j^{i2} I_{m_{ij}}$ où $\gamma_j^i \in \mathbb{R}^+$.

— $F_j^i \sim \mathcal{GP}(\eta_j^i, k_j^i(.,.|\theta_j^i))$ est une fonction issue d'un GP de moyenne constante η_j^i et de fonction de covariance k_j^i paramétrisée par θ_j^i.

On note $\mathcal{P} = \left((\eta_j^i, \theta_j^i, \gamma_j^i)_{i\in[|1,N|],j\in[|1,D|]}, \alpha, \beta, \sigma^2\right) \in P$ est le jeu de paramètres du modèle **(3)**, et on déduit aisément de la propriété **1.** le résultat suivant.

Propriété 2. *Soit $\mathcal{P} \in P$. $\forall i \in [|1,N|]$ la loi jointe de $(Y_i, X_1^i, ..., X_D^i)$ est normale d'espérance :*

$$\mathbb{E}(Y_i) = Z_i^T \alpha + \frac{1}{T_i} \sum_{j=1}^{D} \int_0^{T_i} \eta_j^i \beta_j(s, T_i) ds$$

$$\mathbb{E}(X_j^i) = (\eta_j^i, ..., \eta_j^i)^T \ , \ \forall j \in [|1,D|]$$

et de matrice de variance-covariance :

$$V(Y_i, X^i) = \begin{pmatrix} \sigma^2 + \frac{1}{T_i^2} \sum_j \int_{[0,T_i]^2} k_j^i \beta_j \beta_j & \frac{1}{T_i} \int_0^{T_i} K_*^{i1T} \beta & \cdots & \\ \frac{1}{T_i} \int_0^{T_i} \beta_1 K_*^{i1} & K_1^i + \gamma_1^{i2} I_{m_{i1}} & 0 & \\ \vdots & 0 & \ddots & 0 \\ \frac{1}{T_i} \int_0^{T_i} \beta_D K_*^{iD} & & 0 & K_D^i + \gamma_D^{i2} I_{m_{iD}} \end{pmatrix}$$

Où : $\forall i \in [|1,N|], j \in [|1,D|]$,

$$K_j^i = \begin{pmatrix} k_j^i(t_{j1}^i, t_{j1}^i) & & \\ \vdots & \ddots & \\ k_j^i(t_{jm_{ij}}^i, t_{j1}^i) & \cdots & k_j^i(t_{jm_{ij}}^i, t_{jm_{ij}}^i) \end{pmatrix} \text{ et } \forall s \in \mathbb{R},\ K_*^{ij}(s) = (k(s, t_{j1}^i), ..., k(s, t_{jm_{ij}}^i))^T.$$

On peut à présent utiliser la méthode du maximum de vraisemblance pour obtenir une estimation des paramètres $\mathcal{P}$. On maximise la vraisemblance jointe globale, qui est le produit des vraisemblances jointes locales des N observations indépendantes. Cependant, la matrice de covariance de l'observation i est carrée de taille $1+\sum_j m_{ij}$. Son inversion directe serait coûteuse en calculs lorsque le nombre de variables fonctionnelle et la densité d'échantillonnage augmente. Cela dit, nous pouvons décomposer simplement la vraisemblance jointe du patient i en utilisant l'indépendance des D variables fonctionnelles entre elles, c'est à dire $\forall i \in [|1,N|]$:

$$p(Y_i, X_1^i, ..., X_D^i | \mathcal{P}) \ = \ p(Y_i | X_1^i, ..., X_D^i, \mathcal{P}) \prod_{j=1}^{D} p(X_j^i | \mathcal{P})$$

Une simple extension du corollaire de la partie précédente, dans le cas où F est de moyenne constante et de plusieurs variables fonctionnelles, nous permet d'exprimer la vraisemblance de Y_i conditionnellement aux X_j^i. En effet, la vraisemblance conditionnelle est gaussienne d'espérance et variance :

$$
\begin{aligned}
\mathbb{E}(Y_i|X^i) &:= \mathbb{E}(Y_i|X_1^i(t_1^i),...,X_D^i(t_D^i)) \\
&= Z_i^T\alpha + \tfrac{1}{T_i}\sum_{j=1}^{D}\int_0^{T_i}\mathbb{E}(F_j^i(s)|X_j^i)\,\beta_j(s,T)ds \\
&= Z_i^T\alpha + \tfrac{1}{T_i}\sum_{j=1}^{D}\int_0^{T_i}\left(\eta_j^i + K_*^{ijT}(s).(K_j^i+\gamma^2 I_{m_{ij}})^{-1}.(X_j^i(t_j^i)-\eta_j^i)\right)\beta_j(s,T)ds \\
\mathbb{V}(Y_i|X^i) &:= \mathbb{V}(Y_i|X_1^i(t_1^i),...,X_D^i(t_D^i)) \\
&= \sigma^2 + \tfrac{1}{T_i^2}\sum_{j=1}^{D}\int_{[0,T_i]^2}\mathbb{C}ov\left(F_j^i(s),F_j^i(t)|X_j^i\right)\beta_j(s,T)\beta_j(t,T)dsdt \\
&= \sigma^2 + \tfrac{1}{T_i^2}\sum_{j=1}^{D}\int_{[0,T_i]^2}\left(\,k_j^i(s,t|\theta_j^i) - K_*^{ijT}(s).(K_j^i+\gamma^2 I_{m_{ij}})^{-1}.K_*^{ij}(t)\,\right)\beta_j(s,T)\beta_j(t,T)dsdt
\end{aligned}
$$

On en déduit, à partir de la densité d'un vecteur gaussien, la **log-vraisemblance jointe négative** :

$$
-\log(p(Y,X|\mathcal{P})) = -\sum_{i=1}^{N}\left[\log(p(Y_i|X^i,\mathcal{P})) + \sum_{j=1}^{D}\log(p(X_j^i|\mathcal{P}))\,\right] =
$$

$$
\frac{N}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{N}\left[\,\frac{(Y_i-\mathbb{E}(Y_i|X^i))^2}{\mathbb{V}(Y_i|X^i)} + \log\mathbb{V}(Y_i|X^i)\;+\right.
$$

$$
\left.\sum_{j=1}^{D}\left(X_j^i(t_j^i)-\eta_j^i\right)^T.\Psi_j^i.\left(X_j^i(t_j^i)-\eta_j^i\right) + \log(\det(\Psi_j^{i-1})) + m_{ij}\,\log(2\pi)\,\right]
$$

Où $\Psi_j^i := \left(K_j^i+\gamma_j^{i2}I_{m_{ij}}\right)^{-1}$, $\forall i\in[|1,N|], j\in[|1,D|]$ est la matrice de précision de X_j^i.
On simplifie cette expression avec la forme que l'on notera $\mathcal{L}(Y,X|\mathcal{P})$ en mettant de côté les termes qui ne dépendent d'aucun paramètre et sont donc indifférents à l'optimisation.

$$
\begin{aligned}
&\mathcal{L}(Y,X|\mathcal{P}) = \sum_{i=1}^{N}\left[\mathcal{L}(Y_i|X^i,\mathcal{P}) + \sum_{j=1}^{D}\mathcal{L}(X_j^i|\mathcal{P})\,\right] \\
&= \sum_{i=1}^{N}\left[\,\frac{(Y_i-\mathbb{E}(Y_i|X^i))^2}{\mathbb{V}(Y_i|X^i)} + \log\mathbb{V}(Y_i|X^i)\;\;+\sum_{j=1}^{D}\left(X_j^i(t_j^i)-\eta_j^i\right)^T.\Psi_j^i.\left(X_j^i(t_j^i)-\eta_j^i\right) + \log(\det(\Psi_j^{i-1}))\,\right]
\end{aligned}
$$

La prédiction $\mathbb{E}(Y|X_1,...,X_D)$ de Y, écrite plus haut, dépend des espérances des densités prédictives $\mathbb{E}(F_j^i|X_j^i)$ de chaque variable fonctionnelle. Ainsi, les erreurs de prédictions sont influencées par les paramètres de chacune des D variables fonctionnelles. Le poids de chaque erreur $Y_i - \mathbb{E}(Y_i|X^i)$ est pondéré par la variance de la prédiction $\mathbb{V}(Y_i|X_i)$. Cette dernière est modulée par les variances de chacune des D variables fonctionnelles latentes. Ainsi, on comprend en quoi la maximisation de la vraisemblance de Y_i n'est pas dissociable de celle des X_j^i dans notre modèle hiérarchique.

Remarque : Pondération par l'incertitude de la trajectoire fonctionnelle. Un intérêt de ce modèle est que la variance $\mathbb{V}(Y|X_1,...,X_D)$ est modulée en fonction du nombre de temps d'observations de chaque variable fonctionnelle. En effet, si les temps d'observations sont répartis uniformément sur $[0,T]$, $K_*(s)(K+\gamma^2 I_m)^{-1}K_*(t) \underset{m\to\infty}{\to} k(s,t)$, d'où $\mathbb{V}(Y|X) \underset{m\to\infty}{\to} 0$. Ainsi, l'incertitude sur la trajectoire fonctionnelle de F sera transmise dans la variance de $Y|X_1,...,X_D$ pour tout nouveau patient dont on a observé les signes vitaux au cours du temps. Cela va nous permettre de calculer des probabilités de ré-hospitalisation individuelles à une durée donnée.

5 Méthodes d'inférence

5.1 Choix d'une base de fonctions pour le paramètre fonctionnel

le paramètre $\beta \in \mathbb{R}^{\mathbb{R}^2}$ est un élément dans une espace de dimension infinie que nous voudrions estimer avec un nombre fini de données. Comme le dit l'expression populaire, sans hypothèse supplémentaire, autant chercher une aiguille dans une botte de foin. L'approche que nous avons choisi est celle de la régression fonctionnelle linéaire telle que présentée dans Ramsay [2006]. Il s'agit de modéliser β par une combinaison linéaire de fonctions d'une famille que nous notons par le vecteur $\phi = (\phi_1, ..., \phi_{K_\beta})^T$, tel que pour $b \in \mathbb{R}^{K_\beta}$ le paramètre associé à β et $\forall s, T \in \mathbb{R}^2, \beta(s,T) = \phi^T(s,T)b$. Nous avons choisi un format très simple pour la base fonctionnelle, afin de faciliter une implémentation autonome en python. Pour un degré d donné, on utilise la base polynomiale unidimensionnelle $(1, s, ..., s^{d-1})$ de degré d et on constitue une base de fonctions bidimensionnelles avec l'ensemble des produits de monômes. Cela donne la famille de fonctions $((s,t) \to s^i t^j)_{i,j \in [|0, K_\beta - 1]}$ qui est libre dans l'espace $\mathbb{R}^{\mathbb{R}^2}$.

5.2 Identifiabilité

Rappelons que le modèle d'une variable aléatoire X fondé sur un jeu de paramètres $p \in P$ est identifiable si et seulement si, $\forall p, p' \in P$:

$$Loi(X|p) = Loi(X|p') \Rightarrow p = p'$$

Modèle sur X Intéressons nous d'abord à l'identifiabilité de la modélisation choisie pour le signal fonctionnel, à savoir le modèle de régression sur processus gaussien **(1)** de la partie **4**. Pour rappel, on note $t = (t_1, ..., t_m)^T$ les instants d'observations du signal, où $m \geq 1$, et $X = X(t) = (X(t_1), ..., X(t_m))^T$ le vecteur aléatoire des observations en ces instants. On modélise X par :

$$X(t) = F(t) + \zeta \quad \textbf{(1)}$$

où $\zeta \sim \mathcal{N}(0_m, \gamma^2 I_m)$, F est un processus gaussien de fonction moyenne constante et égale à $\eta \in \mathbb{R}$ et de fonction de covariance k sur $\mathbb{R}^2$ de paramètres $\theta = (\theta_1, ..., \theta_d) \in \mathbb{R}^d$. Lorsque k est une fonction de la distance euclidienne entre les instants d'observation, on dit qu'elle est stationnaire. C'est le cas pour la plupart des fonctions de covariances utilisées et on prendra cette hypothèse de stationnarité.

Il existe des cas de fonctions de covariances, jeux de paramètres et de données pour lesquels le modèle suivant n'est pas identifiable pour plusieurs raisons dont nous présentons quelques exemples simples. Tout d'abord, les données collectées, même en nombre, peuvent souvent dissimuler un effet de covariance. Prenons l'exemple du noyau périodique dont l'expression est :

$$k(r) = \sigma^2 e^{\frac{-2}{l^2} sin^2(r\pi/p)}$$

où r est la distance absolue entre les points, σ le paramètre d'échelle, l la portée et p la périodicité du signal. Lorsque les points sont espacés selon une distance égale à a période, le paramètre de période n'est pas identifiable puisque la matrice de variance covariance devient une matrice de σ^2. Un autre cas de figure peut se présenter à la limite des valeurs d'un paramètre. Le cas simple du noyau exponentiel quadratique peu témoigner de ce phénomène. son expression est

$$k(r) = \sigma^2 e^{\frac{-r^2}{l^2}}$$

où on retrouve les paramètre d'échelle et de portée. Lorsque le σ devient petit, même avec beaucoup de points, l'effet de la portées sur la loi de X devient faible, la valeur de l devient indiscernable. Par ailleurs, lorsque l est grand la matrice de variance-covariance de X est proche d'une matrice de σ.

En fait, cette propriété du modèle de régression sur processus gaussien n'est pas fondamentalement un problème vis à vis de l'objectif visé. Tout d'abord, nous sommes intéressés ici par l'estimation d'un paramètre fonctionnel sur un jeu de données fonctionnel observé sur lequel on réalise de nombreuses régression sur processus gaussiens. Ainsi, les cas de non-identifiabilité sont plutôt rares dans l'échantillon et affectent peu l'estimation du paramètre fonctionnel. En deuxième lieu, la confusion des lois est souvent liée à une multiplicité des interprétations que l'on peut faire de la série temporelle observée quant à sa courbe sous-jacente, comme le reflète l'exemple précédent du noyau périodique, et les paramètres non-identifiés n'affecteront pas forcément la forme sur le domaine d'étude. Cependant, il semble plus prudent de choisir une paramétrisation de la fonction de covariance dont le nombre de paramètre n'excède pas le nombre minimal de points contenues dans les variables fonctionnelles observées. Cette règle est importante dans notre cas de faible densité d'échantillonnage, et on veille à garder des noyaux simples.

Modèles sur $Y|X$ A présent, on va montrer l'identifiabilité du modèle du label Y connaissant la loi des variables fonctionnelles. Pour cela, on simplifie un peu le modèle **(3)** en considérant, sans perte de généralité, une seule variable fonctionnelle ($D = 1$) et en retirant le terme lié aux variables scalaires Z_i. De plus, on suppose que les paramètres liés aux variables fonctionnelles $(\gamma_i, \theta_i)_{i\in[|1,N|]}$ sont fixés. On pose $F_1, ..., F_N$ des processus gaussiens de fonction moyenne connues $f_1, ..., f_N$. On a alors $\forall_i \in [|1, N|]$:

$$Y_i = \frac{1}{T_i}\int_0^{T_i} F_i(s)\beta(s, T_i)ds + \epsilon_i$$

Les $(Y_i)_{i\in[|1,N|]}$ sont indépendants et rappelons que, d'après la propriété de la partie **4**, leur loi est gaussienne Y connue. On s'intéresse à l'identifiabilité du paramètre $b = (b_1, ..., b_{K_\beta})^T$ dans ce modèle.

On note $l := \begin{pmatrix} \frac{1}{T_1}\int_0^{T_1} f_1(s)\Phi^T(s, T_1)ds \\ \vdots \\ \frac{1}{T_N}\int_0^{T_N} f_N(s)\Phi^T(s, T_N)ds \end{pmatrix}$

Propriété :
Si l est de rang égal à K_β, alors le modèle précédent est identifiable pour b.

Preuve :

Montrons que pour $b, b' \in \mathbb{R}^{K_\beta}$, $\mathcal{L}(Y|b) = \mathcal{L}(Y|b') \Rightarrow b = b'$.
Soit $b, b' \in \mathbb{R}^{K_\beta}$, supposons $\mathcal{L}(Y|b) = \mathcal{L}(Y|b')$. Alors, on a $E(Y|b) = E(Y|b') \Rightarrow lb = lb' \Leftrightarrow l(b - b') = 0_N$. Par conséquent, comme l est de rang égal à la dimension de $b - b'$, ce système linéaire admet une unique solution qui est $b - b' = 0_N \Leftrightarrow b = b'$. □

On étend par un raisonnement similaire cette propriété à l'ajout du terme $Z_i^T\alpha$ dans l'expression de Y_i, en précisant que $N \geq K_\beta + p$ et que la matrice de design doit être de rang $K_\beta + p$.

5.3 Propriétés analytiques et existence d'un minimum

Pour simplifier notations et calculs, et sans perte de généralité des résultats, on ne considérera ici qu'une seule variable fonctionnelle dans le modèle **(3)**, on laisse ainsi de côté la notation de l'indice associé j.

Notons $l = (l_1, ..., l_N)^T$ où $\forall i \in [|1, N|], l_i^T := \frac{1}{T_i}\int_0^{T_i} \mathbb{E}(F^i(s)|X^i)\phi^T(s, T_i)ds$ est la $i$ème ligne de l.
On a $\mathbb{E}(Y_i|X^i) = l_i^T b + Z_i^T\alpha$.
et $\forall i \in [|1, N|]$, $V^i := \frac{1}{T_i^2}\int_{[0,T_i]^2} \mathbb{C}ov(F^i(s), F^i(t)|X^i)\Phi(s, T_i)\Phi^T(t, T_i)dsdt$, c'est à dire que d'après ce

qui précède, $\mathbb{V}(Y_i|X^i, b) = \sigma^2 + b^T V^i b$.

Nous allons étudier la convexité de la fonction $\mathcal{P} \to \mathcal{L}(Y|X, \mathcal{P})$, en tant que fonction des paramètres α et b. Son expression est la suivante :

$$\mathcal{L}(Y|X, \alpha, b) = \sum_{i=1}^{N} \frac{(Y_i - l_i^T b - Z_i^T \alpha)^2}{\sigma^2 + b^T V^i b} + \log(\sigma^2 + b^T V^i b)$$

Convexité par rapport à α. Considérons la fonction $f_b : \alpha \to \mathcal{L}(Y|X, b, \alpha)$ définie sur $\mathbb{R}^p$.
$\mathbb{V}(Y_i|X^i)$ est indépendant de α, donc constant lorsque ce paramètre varie, et toujours strictement positif car $\sigma^2 > 0$. Ainsi on peut mettre de côté ces facteurs dans l'expression de f_b, quitte à diviser les lignes de Z et les éléments de $Y - lb$ par les $\sqrt{\mathbb{V}(Y_i|X^i)}$. On retrouve alors la forme :

$$\begin{aligned} f_b(\alpha) &= \langle Z\alpha, Z\alpha \rangle - 2 \langle Y - lb, Z\alpha \rangle + \text{Constante} \\ &= \langle Z^T Z\alpha, \alpha \rangle - 2 \langle Y - lb, Z\alpha \rangle + \text{Constante} \end{aligned}$$

La matrice $Z^T Z \in \mathcal{M}_{p,p}$ est symétrique semi-définie positive, car $\forall \alpha \in \mathbb{R}^p$ non nul, on a $\alpha^T Z^T Z \alpha = (Z\alpha)^2 \geq 0$. Cela implique que f_b est une forme quadratique, éventuellement dégénérée, convexe. Elle admet alors au moins un minimum $\mathbb{R}^p$.

Remarque : Si on se place dans un contexte de petite dimension du point de vue des variables scalaires Z_i, c'est à dire que $N \geq p$. On peut faire l'hypothèse que la matrice de design Z est de rang p. Dans ce cas $Z^T Z$ devient définie positive. On peut alors affirmer que $\forall b \in \mathbb{R}^{K_\beta}$, f_b est une fonctionnelle quadratique fortement convexe sur $\mathbb{R}^p$ et cela implique l'unicité du minimum de f_b.

Non-convexité et coercivité par rapport à b. A présent, étudions le comportement par rapport à b, c'est à dire la fonction $f_\alpha : b \to \mathcal{L}(Y|X, b, \alpha)$.

On suppose que $N \geq K_\beta$, que l et $V^i, \forall i$ sont de rang K_β. Or, on sait déjà que ces matrices sont semi-définies positives car $\forall b \in \mathbb{R}^{K_\beta}$, $b^T V^i b = \mathbb{V}\left(\int_0^{T_i} \beta(s, T_i) F^i(s) ds | X^i\right) \geq 0$ et $b^T l^T l b = \langle lb, lb \rangle \geq 0$ par positivité du produit scalaire. En ajoutant la condition qu'elles soient de rang K_β, elles deviennent définies positives. Cette propriété aboutit aux résultat suivant.

Propriété :
Supposons que $\forall i \in [|1, N|], V^i$ soit définie positive. Alors, $\forall \alpha \in \mathbb{R}^{K_\beta}$, f_α est non-convexe et coercive.

Preuve :
Coercivité : V^i est définie positive, alors $b^T V^i b = \langle V_i b, b \rangle$ est une forme quadratique non-dégénérée, donc $b^T V^i b \underset{||b|| \to \infty}{\to} +\infty$, et comme $\log(x) \underset{x \to +\infty}{\to} +\infty$, alors, par composition des limites : $\log(b^T V^i b) \underset{||b|| \to \infty}{\to} +\infty$. De plus la fonction est positive en l'infini car le plus haut terme du polynôme au numérateur du terme de gauche est quadratique :

$$\frac{(Y_i - l_i^T b - Z_i^T \alpha)^2}{\sigma^2 + b^T V^i b} \underset{||b|| \to +\infty}{\approx} \frac{(l_i^T b)^2}{b^T V^i b} \geq 0$$

Par conséquent, par somme d'un terme positif et d'un terme qui tend vers l'infini, f_α est coercive.
Non-convexité : Écrivons le gradient du $i$ème terme :

$$\frac{\partial f_\alpha^i}{\partial b} = \frac{-2l_i(Y_i - Z_i^T \alpha - l_i^T b)(b^T V^i b + \sigma^2) - 2V^i b(Y_i - Z_i^T \alpha - l_i^T b)^2}{(b^T V^i b + \sigma^2)^2} + \frac{2V^i b}{b^T V^i b + \sigma^2}$$

Soit $k \in [|1, K_\beta|]$, montrons que le gradient relativement à b_k s'annule lorsque $b_k \to \infty$. C'est évident pour le deuxième terme issu de la dérivée du logarithme en comparant les termes de plus haut

rangs en b_k qui sont $2V^i_{kk}b_k$ au numérateur et $V^i_{kk}b_k^2$ au dénominateur. Ce terme est donc équivalent à $O(2/b_k)$ lorsque $b_k \to \infty$ et s'annule.
En ce qui concerne le numérateur du terme de gauche, le terme d'ordre 3 de ce polynome s'annule, il reste un polynôme du second degré de terme quadratique : $2l_{ik}\left((Y_i - Z_i^T\alpha)V^i_{kk} + l_{ik}(\sum_{j\neq k} V^i_{jk}b_j)\right)b_k^2$. Or, le dénominateur est un polynôme de degré 4. Alors, ce terme est équivalent à $O\left(\frac{1}{b_k^2}\right)$ en l'infini et s'annule également.
On a alors $\forall i \in [|1, N|]$, $\frac{\partial f^i_\alpha}{\partial b_k} \to 0 \Rightarrow \frac{\partial f_\alpha}{\partial b_k} \to 0$. Comme c'est vrai $\forall k \in [|1, K_\beta|]$, on a $\frac{\partial f_\alpha}{\partial b} \to 0_{K_\beta}$. Alors, comme la fonction est non-constante, elle n'est pas convexe. En effet, au moins l'une de ses composantes atteint sur $\mathbb{R}$ une valeur non-nulle, et cette composante doit alors décroître sur un intervalle de $\mathbb{R}$.

□

Remarque : On peut noter que la pente est faible en l'infini, car c'est le terme $\log(b^TV^ib)$ qui entraîne la coercivité. En effet, on peut montrer que l'autre terme est borné, $\forall i \in [|1, N|], V^i$ étant définie positive, $b \to b^TV^ib$ est fortement convexe, il existe donc $C^i > 0/\ \forall b \in \mathbb{R}^{K_\beta}, b^TV^ib \geq C||b||^2$. De plus, $(Y_i - Z_i^T\alpha - l_i^Tb)^2 \underset{||b||\to+\infty}{\approx} (l_i^Tb)^2 \leq ||l_i||^2||b||^2$. Ainsi, on peut majorer le terme de gauche par une constante positive :

$$\frac{(Y_i - l_i^Tb - Z_i^T\alpha)^2}{\sigma^2 + b^TV^ib} \underset{||b||\to+\infty}{\approx} \frac{b^Tl_il_i^Tb}{b^TV^ib} \leq \frac{||l^i||^2||b||^2}{C^i||b||^2} = \frac{||l^i||^2}{C^i}$$

$$\Rightarrow \exists C_1, ..., C_N > 0\ /\ 0 \underset{||b||\to+\infty}{\leq} \frac{(Y_i - l_i^Tb - Z_i^T\alpha)^2}{\sigma^2 + b^TV^ib} \underset{||b||\to+\infty}{\leq} \sum_{i=1}^{N} \frac{||l^i||^2}{C^i}$$

Comme le terme est continu et positif en l'infini , il est borné. On en déduit qu'il rejoint une asymptote constante en l'infini.

Remarque : Le fait que les matrices V_i soient définies positives est important, car si ce n'est pas le cas on n'obtient pas forcément la coercivité de f_α. Soit $i \in [|1, N|]$, si V_i n'est pas définie positive, comme elle est symétrique semi-définie positive, elle est de rang inférieur strictement à K_β, et il existe donc un vecteur de coefficients réels $C = (c_1, ..., c_{K_\beta})$ non nul tel que les colonnes de V^i soient linéairement liées par $\sum_{k=1}^{K_\beta} c_kV^i_{.k} = 0_{K_\beta}$. Alors f_α ne peut être coercive car quelque soit $\epsilon > 0$, il n'existe aucun $\rho \geq 0/\forall b \in \mathbb{R}^{K_\beta}, ||b|| \geq \rho \Rightarrow b^TV^ib > \epsilon$. En effet par l'absurde, soit $\epsilon > 0$, supposons qu'un ρ remplissant l'assertion précédente existe. Alors, en prenant $b = \frac{\rho}{||C||}(c_1, ..., c_{K_\beta})$ on a bien $||b|| = \rho \geq \rho$ et par ailleurs $V^ib = 0_{K_\beta} \Rightarrow b^TV^ib = 0$, il y a donc contradiction. Dans ce cas, le terme logarithmique ne tend pas nécessairement vers l'infini en l'infini dans tout direction, et on ne peut rien dire de simple quant à la coercivité de l'autre terme.

On déduit aisément des résultats qui précèdent que $\alpha, b \to \mathcal{L}(Y|X, \alpha, b)$ est continue et coercive sur $\mathbb{R}^{p+K_\beta}$. Alors, d'après le théorême d'existence d'un minimum, il existe au moins un minimum global à $\mathcal{L}(Y|X)$ sur $\mathbb{R}^{p+K_\beta}$. Ce résultat n'exclue pas cependant l'existence de minima locaux.

Comportement de $\mathcal{L}(Y_i, X^i)$ par rapport à θ^i, γ^i. Il est montré dans Rasmussen [2006] que la log-vraisemblance négative gaussienne des données fonctionnelles, que nous appelons $\forall i \in [|1, N|], \mathcal{L}(X^i|\theta^i, \gamma^i, \eta^i)$ n'est pas convexe dans le cas général par rapport aux paramètres $(\gamma^i, \theta^i)_{i\in[|1,N|]}$ de la fonction de covariance du processus gaussien, et qu'elle peut admettre des minima locaux. Cette propriété s'étend à $\mathcal{L}(Y_i, X^i)$ Cette information est importante à retenir pour l'inférence simultanée de tous les paramètres du modèle **(3)**, mais nous ne montrerons par ce résultat ici.

5.4 Un critère pénalisé

La minimisation directe du critère précédent, qui revient à maximiser la fonction de vraisemblance relativement aux paramètres $\mathcal{P}$, n'est pas la méthode la plus adaptée pour obtenir une estimation fiable et robuste des paramètres dans notre contexte où $Card(\mathcal{P})$ est du même ordre que N. Cela est développé en détails dans Ramsay [2006]. Nous nous contenterons ici de fournir une explication intuitive du problème. Considérons les paramètres liés à la régression sur le label censuré : $\alpha \in \mathbb{R}^d$ est de dimension relativement faible par rapport à N. En revanche, $(b_j)_{j_i n[|1,D|]}$ est de dimension $K_\beta * D$. Or, si l'on souhaite disposer d'une flexibilité raisonnable sur la modélisation des fonctions $(\beta_j)_{j\in[|1,D|]}$, il faut utiliser une base fonctionnelle suffisamment étendue où on sera confronté à un problème de biais dans l'estimation des $\hat{\beta}_j$. Nous sommes d'autant plus exposé à ce problème car nos fonctions de base sont bidimensionnelles, il nous faut donc choisir K_β relativement grand, et donc le nombre de paramètres du modèle de régression sur Y^c est de l'ordre de $K_\beta D$. Par exemple, si on prend $K_\beta = 20^2 = 400$, pour des données hospitalières typiques contenant 5 variables temporelles explicatives et 500 patients, on a déjà quatre fois plus de paramètres que d'observations, ce qui annonce un risque de sur-ajustement des β_j aux données. On veut alors imposer des contraintes de régularité sur ces fonctions, selon l'approche de la régression fonctionnelle pénalisée (Ramsay [2006],Goldsmith *et al.* [2012]). Une pénalité qui devrait nous permettre d'utiliser une grande base de fonctions pour modéliser les β_j sans pour autant sur-ajuster notre modèle aux données est la pénalité de sinuosité telle qu'elle est utilisée dans Wood [2003] :

$$Pen(\beta) = J_{22}(\beta) = \sum_{j=1}^{D} \int_0^{Tmax} \int_0^t \sum_{\nu_1+\nu_2=2} \frac{2}{\nu_1!\,\nu_1!} \left(\frac{\partial^2 \beta_j(s,t)}{\partial s^{\nu_1} \partial t^{\nu_2}} \right)^2 ds\, dt$$

La pénalité peut se développer en $Pen(\beta) = \sum_{j=1}^{D} b_j^T (I_s + 2I_c + I_t) b_j$, où I_s, I_c, I_t sont les matrices d'intégrales produit des dérivées secondes de la base fonctionnelle $\phi = (\phi_1, ..., \phi_{K_\beta})$, c'est à dire :

$I_s := \int_0^{Tmax} \int_0^t \frac{\partial^2 \phi(s,t)}{\partial s^2} . \frac{\partial^2 \phi(s,t)}{\partial s^2}^T ds\, dt$

$I_t := \int_0^{Tmax} \int_0^t \frac{\partial^2 \phi(s,t)}{\partial t^2} . \frac{\partial^2 \phi(s,t)}{\partial t^2}^T ds\, dt$

$I_c := \int_0^{Tmax} \int_0^t \frac{\partial^2 \phi(s,t)}{\partial s \partial t} . \frac{\partial^2 \phi(s,t)}{\partial s \partial t}^T ds\, dt$

Calcul : On a une forme identique quel que soit le terme de dérivée partielle de second ordre. $\forall j \in [|1, D|]$:

$$\int_O^{Tmax} \int_0^t \left(\frac{\partial^2 \beta_j(s,t)}{\partial . \partial .} \right)^2 dsdt$$

$$= \int_O^{Tmax} \int_0^t \left(\frac{\partial^2 \phi(s,t)^T}{\partial . \partial .} b_j \right)^2 dsdt$$

$$= b_j^T \left(\int_O^{Tmax} \int_0^t \frac{\partial \phi(s,t)}{\partial . \partial .} \frac{\partial \phi(s,t)^T}{\partial . \partial .} dsdt \right) b_j = b_j^T \; I_. \; b_j$$

De plus, comme d'après le théorème de Schwartz, les dérivées partielles de second ordre croisées sont égales, on a : $J_{22}(\beta) = \sum_{j=1}^{D} \sum_{\nu_1+\nu_2=2} \frac{2}{\nu_1!\,\nu_1!} \int_0^{Tmax} \int_0^t \left(\frac{\partial^2 \beta_j(s,t)}{\partial s^{\nu_1} \partial t^{\nu_2}} \right)^2 ds\, dt =$

$\sum_{j=1}^{D} b_j^T I_s b_j + 2 b_j^T I_c b_j + b_j^T I_t b_j = \sum_{j=1}^{D} b_j^T (I_s + 2I_c + I_t) b_j \quad \square$

Les propriétés de la pénalité choisie est assez similaire avec la pénalité Ridge (CITATION) car il s'agit d'une somme quadratique des paramètres de régression $(b_{jk})_{k\in[|1,K_\beta|],j\in[|1,D|]}$, mais elle ajoute une pondération naturelle de chaque coefficient de b par la courbure de la fonction de base associée. On aurait aussi pu envisager d'autres forme de pénalisation ... **(à détailler ?)**

Pour les paramètres du modèle **(3)**, on considérera donc l'estimateur :

$$\hat{\mathcal{P}} = \underset{\mathcal{P}\in P}{argmin}\left\{\mathcal{L}^p_\lambda(Y,X|\mathcal{P})\right\} = \underset{\mathcal{P}\in P}{argmin}\left\{\mathcal{L}(Y,X|\mathcal{P}) + \lambda Pen(\mathcal{P})\right\}$$

Où $\lambda \in \mathbb{R}^+$ est un hyper-paramètre du modèle qui est fixé avant l'optimisation.

Remarque : La régularisation procure un avantage notoire pour l'optimisation. En effet, la sous-partie précédente a montré que le critère $\mathcal{L}(Y|X,\alpha,b)$ était coercif relativement à b, cependant, la pente s'annule en l'infini ce qui veut dire que la convergence vers un minimum local par un algorithme de gradient peut être très lente si le point de départ est "éloigné". Ainsi, cette pénalité quadratique accélère nécessairement la convergence en rendant la pente infinie en l'infini.

5.5 Méthode du maximum de vraisemblance en deux étapes

Principe. Plaçons nous au niveau d'un individu $i \in [|1,N|]$, jusqu'à présent nous avons parlé de la loi jointe du vecteur aléatoire X^i et de la variable Y^i à partir de laquelle nous avons produit un critère basé sur la vraisemblance qu'il s'agirait de minimiser, selon le principe du maximum de vraisemblance, pour obtenir une estimation des paramètres du modèle. Cela dit, cette approche est relativement ambitieuse au vu de la dimmensionalité de la fonction, et de la complexité de son évaluation. Aussi, nous voulons recourir à plusieurs méthodes d'inférence et être à même comparer leurs performances en terme d'estimation et de temps de calcul pour décider quelle serait la plus adaptée à nos applications. Par ailleurs, de nombreuses méthodes d'inférence optimisées existent déjà pour la régression sur processus gaussien et il est bien pratique de les exploiter pour obtenir une inférence des courbes fonctionnelles, que nous pouvons mobiliser dans un deuxième temps et maximiser la vraisemblance conditionnelle de $Y|X$ relativement aux paramètres généraux α, β. Décrivons en détails cette méthode.

1. La première étape consiste en une régression sur processus gaussien pour chaque variable fonctionnelle $i \in [|1,N|]$ selon le modèle **(1)** présenté dans la partie **4** qui consiste à déterminer les paramètres de la fonction de covariance de processus gaussien θ^i, la moyenne η^i et la variance de l'erreur gaussienne indépendante γ^i. La méthode est présentée dans l'appendice.

2. Une fois qu'on a ces paramètres en mains, pour des paramètres α, b on connaît la loi gaussienne des $Y^i|X^i$. La deuxième étape consiste alors à calculer les estimateurs $\hat{\alpha}, \hat{b}$ tels que :

$$\hat{\alpha}, \hat{b} = \underset{\alpha\in\mathbb{R}^p, b\in\mathbb{R}^{K_\beta}}{\operatorname{argmin}} \mathcal{L}\left(Y|X,\alpha,b\right) + \lambda Pen(\beta)$$

Où $\mathcal{L}$ dépend implicitement de $(\theta_i,\eta_i,\gamma_i^2)_{i\in[|1,N|]}$ et où λ et $\sigma^2 > 0$ sont fixés à priori.

Pour y parvenir, on définit une routine d'optimisation simple basée sur un algorithme de descente de gradient. On stoppe l'algorithme lorsque la norme du gradient descend en dessous d'un seuil fixé. Le calcul de la hessienne pourrait permettre d'utiliser des algorithmes plus rapides, la complexité des formules fut dissuasive, et aurait pu mener à de nombreuses erreurs d'implémentation. On suppose que le nombre d'observation est suffisant pour assurer la matrice des variables scalaires Z soit de rang p et que le critère soit fortement convexe par rapport à α. En revanche, au vu de l'existence probable de minima locaux relativement au paramètre b, le lancement d'un seul algorithme de descente de gradient jusqu'à convergence n'est pas forcément suffisante pour atteindre un minimum global de la

fonction. On choisit donc de réaliser un nombre, fixé à l'avance, d'initialisations aléatoires du paramètre b et de garder l'estimateur minimal pour la fonction objectif.

Gradient. Rappelons les quelques notations employées dans la partie précédente. On note la matrice de design fonctionnel $l := (l_1, ..., l_N)^T$ où $\forall i \in [|1, N|]$:

$$l_i^T = \left(\frac{1}{T_i} \int_{[0,T_i]} \phi_1(s, T_i) \mathbb{E}(F^i(s)|X^i) ds, ..., \frac{1}{T_i} \int_{[0,T_i]} \phi_{K_\beta}(s, T_i) \mathbb{E}(F^i(s)|X^i) ds \right)$$

Où l'espérance conditionnelle $\mathbb{E}(F^i(s)|X^i)$ est connue et explicite pour tout s, car la première étape a livré les paramètres fonctionnels $\theta^i, \gamma^i, \eta^i$. Par ailleurs $\forall i \in [|1, N|]$ la matrice de covariance intégrée de i est :

$$\begin{aligned} V^i &= \frac{1}{T_i^2} \int_{[0,T_i]^2} \mathbb{C}^i(s,t) \Phi(s, T_i) \Phi^T(t, T_i) ds dt \\ &= \frac{1}{T_i^2} \int_{[0,T_i]^2} \begin{pmatrix} \mathbb{C}^i(s,t)\Phi_1(s,T_i)\Phi_1(t,T_i) & & \\ \vdots & \ddots & \\ \mathbb{C}^i(s,t)\Phi_{K_\beta}(s,T_i)\Phi_1(t,T_i) & & \mathbb{C}^i(s,t)\Phi_{K_\beta}(s,T_i)\Phi_1(t,T_i) \end{pmatrix} ds dt \quad \text{On} \end{aligned}$$

$$\text{Où } \mathbb{C}^i(s,t) = \mathbb{C}ov(F^i(s), F^i(t)|X^i) = (k^i(s,t) - K^{iT} * (s) \Psi^i K^i * (t))$$

utilise les notations rapides suivantes pour l'espérance et la variance conditionnelle du label :

$$\mathbb{E}_i = \mathbb{E}(Y_i|X^i) = Z_i^T \alpha + l_i^T b$$

$$\mathbb{V}_i = \mathbb{V}(Y_i|X^i) = \sigma^2 + b^T V^i b$$

Écrivons à présent la formule du gradient.

$$\frac{\partial}{\partial \alpha} \mathcal{L}_\lambda^p(Y|X) = -2 \sum_{i=1}^N \frac{\partial \mathbb{E}_i}{\partial \alpha} \frac{Y_i - \mathbb{E}_i}{\mathbb{V}(Y_i|X^i)} = -2 \sum_{i=1}^N Z_i \frac{(Y_i - Z_i^T \alpha - l_i^T b)}{\sigma^2 + b^T V^i b}$$

$$\frac{\partial}{\partial b} \mathcal{L}_\lambda^p(Y|X) = \sum_{i=1}^N -2 \frac{\partial \mathbb{E}_i}{\partial b} \frac{Y_i - \mathbb{E}_i}{\mathbb{V}_i} + \frac{1}{\mathbb{V}_i} \frac{\partial \mathbb{V}_i}{\partial b} \left(1 - \frac{(Y_i - \mathbb{E}_i)^2}{\mathbb{V}_i} \right) + 2\lambda I_p b$$

$$= 2 \sum_{i=1}^N -l_i \frac{(Y_i - Z_i^T \alpha - l_i^T b)}{b^T V^i b + \sigma^2} + V^i b \left(\frac{1}{b^T V^i b + \sigma^2} - \frac{(Y_i - Z_i^T \alpha - l_i^T b)^2}{(b^T V^i b + \sigma^2)^2} \right) + 2\lambda I_p b$$

Technique d'approximation. Pour évaluer le critère et son gradient, nous avons en premier lieu besoin de calculer les matrices l_i et V^i. Nous pouvons soit calculer les différentes intégrales explicitement, soit les approximer. Étant donné que le calcul de l'intégrale dépend du type de noyau de covariance choisi pour la modélisation et que nous voudrions laisser le champs libre sur celui-ci, il est plus pratique d'approximer les intégrales. Le choix s'est porté sur la méthode des trapèzes, aussi classique qu'efficace, basée sur une discrétisation de l'intervalle d'intérêt qui est facile à programmer, rapide et générale quelque soit le noyaux de covariance. On fixe donc un paramètre de finesse J tel

que l'intégrale d'une fonction $f : s \to f(s)$ sur $[O, T]$ sera approximé par

$$\frac{T}{J}\sum_{j=0}^{J} f\left(\frac{jT}{J}\right)\left(1-\frac{1}{2}\delta_{j\in\{0,J\}}\right)$$

Pour l'intégrale en dimension 2 d'une fonction $f : s, t \to f(s, t)$, on compose simplement l'approximation précédente

$$\frac{T^2}{J^2}\sum_{k=0}^{J}\sum_{j=0}^{J} f\left(\frac{jT}{J}, \frac{kT}{J}\right)\left(1-\frac{1}{2}\delta_{j\in\{0,J\}}\right)\left(1-\frac{1}{2}\delta_{k\in\{0,J\}}\right)$$

Algorithme du gradient stochastique. On utilise un algorithme du gradient stochastique à pas linéairement décroissant avec un nombre maximal d'itération $S \in \mathbb{N}$ avec un pas initial $\eta > 0$. A chaque étape, la valeur du gradient est approchée par le gradient de la vraisemblance d'un seul individu choisi uniformément au hasard. Décrivons l'algorithme à une étape $n \in \mathbb{N}$, avec le paramètre $\mathcal{P}_n := (\alpha_n, b_n)^T$:

— Tirage aléatoire de $i \in \{1, ..., N\}$.

— Calcul de $\nabla\mathcal{L}_\lambda^p(Y_i|X^i, \mathcal{P}_n)$.

— Arrêt si $||\nabla\mathcal{L}_\lambda^p(Y_i|X^i, \mathcal{P}_n)|| \leq tol$ où $tol > 0$.

— renvoie de $\mathcal{P}_{n+1} = \mathcal{P}_n - \eta\frac{S-n}{S}\nabla\mathcal{L}_\lambda^p(Y|X, \mathcal{P}_n)$.

5.6 Méthode des moindres carrés pénalisée

On considère toujours que la première étape de la méthode décrite précédemment à été réalisée, à savoir l'estimation des paramètres fonctionnels. Une autre méthode d'inférence que celle du maximum de vraisemblance peut être intéressante à étudier, la méthode des moindres carrés. Rappelons que cette méthode classique vise à minimiser la somme des carrés des écarts de prédiction du modèle, les estimateurs associés sont donc :

$$\hat{\alpha}_{mc}, \hat{b}_{mc} = \underset{\alpha\in\mathbb{R}^p, b\in\mathbb{R}^{K_\beta}}{\operatorname{argmin}} ||Y - Z\alpha - lb||_2^2$$

Dans le modèle **(3)**, la variance de chaque Y_i dépend d'une composante propre issue de la variable fonctionnelle X^i. Il en découle que $Y_1, ..., Y_N$ ne sont pas identiquement distribués. Pour cette raison, l'estimateur des moindres carrés ne correspond pas à celui du maximum de vraisemblance. Cependant, si l et Z sont de rang maximal, il est unique et calculable. Ceci très avantageux car cela balaie les problèmes d'optimisations que nous rencontrerions par la méthode du maximum de vraisemblance, et promet un gain de temps de calcul considérable. Cependant, au vu des ordres de grandeur de K_β et N que nous considérons, il est possible que l soit de rang inférieur à K_β, auquel cas nous ne pouvons estimer les paramètres sans plus d'hypothèse. On remarque alors que si nous ajoutons au critère des moindres carré la régularisation précédente sur le paramètre b et que I_p est de rang K_β, on peut calculer l'estimateur qui minimise ce critère. Pour $\lambda > 0$ il s'écrit, on note $\mathcal{P} = (\alpha^T, b^T)^T$, et on note les matrices carrées $M = Z \oplus b$ et $H = 0_{pp} \oplus I_p$ de dimension $p + K_\beta$. l'estimateur s'écrit alors :

$$\hat{\mathcal{P}}^{mcp} = \underset{\alpha\in\mathbb{R}^p, b\in\mathbb{R}^{K_\beta}}{\operatorname{argmin}} ||Y - Z\alpha - lb||_2^2 + \lambda b^T I_p b = (M^T M + \lambda H)^{-1} M^T Y$$

Nous ne nous étendrons pas sur les propriétés de cet estimateur mais on peut grossièrement remarquer que la pénalité assure que le paramètre fonctionnel estimé ne soit pas trop sinueux ce qui est utile dans le contexte de moyenne dimension où la taille de l'échantillon n'est pas très grande par rapport à la taille de la base fonctionnelle, et favorise le sur-ajustement aux données.

5.7 Méthode jointe

Principe. Revenons sur le modèle-joint à proprement et proposons une procédure d'inférence par le maximum de vraisemblance. Il s'agit ici de trouver les estimateurs vérifiant :

$$\begin{aligned}
\hat{\alpha}, \hat{b}, (\hat{\theta}^i, \hat{\gamma}^i, \hat{\eta}^i)_{i\in[|1,N|]} &= \underset{\alpha,b,(\theta^i,\gamma^i,\eta^i)_{i\in[|1,N|]}}{\operatorname{argmin}} \quad \mathcal{L}(Y|X) + \mathcal{L}(X) + \lambda Pen(\beta) \\
&= \underset{\alpha,b,(\theta^i,\gamma^i,\eta^i)_{i\in[|1,N|]}}{\operatorname{argmin}} \quad \sum_{i=1}^{N} \frac{(Y_i - \mathbb{E}_i)^2}{\mathbb{V}_i} + \log \mathbb{V}_i \\
&\qquad + \sum_{i=1}^{N} \left(X^i - \eta^i\right)^T . \Psi^i . \left(X^i - \eta^i\right) + \log(\det(\Psi^{i-1})) \\
&\qquad + \lambda b^T I_p b
\end{aligned}$$

Tentons d'expliquer l'idée qui sous-tend la méthode d'inférence jointe. Le modèle joint introduit un lien entre la variable Y_i et la variable fonctionnelle X^i via le processus gaussien caché F_i par deux voies que l'on peut observer dans la formule ci-dessus :

- D'une part la trajectoire moyenne de $F_i|X^i$ joue sur l'espérance de Y_i, et la vraisemblance d'une observation de cette variable augmente naturellement lorsqu'on diminue l'écart $Y_i - \mathbb{E}(Y^i|X^i)$. Or, cette trajectoire moyenne ne dépend pas que des points X^i mais aussi de θ^i et η^i.
- D'autre part, la covariance conditionnelle temporelle de $F_i|X^i$ impacte la variance de Y_i et aussi donc la vraisemblance d'une observation. Grossièrement, si $|Y_i - \mathbb{E}(Y^i|X^i)|$ est proche de 0, $p(Y_i|X^i)$ augmentera en faisant diminuer $\mathbb{V}(Y_i|X^i)$, alors que si $|Y_i - \mathbb{E}(Y^i|X^i)|$ est grand, $p(Y_i|X^i)$ augmente lorsque l'on fait croître $\mathbb{V}(Y_i|X^i)$. La covariance conditionnelle dépend de γ^i et θ^i.

Or, ces deux fonctions dépendent des paramètres fonctionnels $\theta^i, \gamma^i,$. Ainsi, l'information de Y_i est réciproquement utilisée pour déterminer les paramètres fonctionnels de F^i. On pourrait dire que quand les paramètres sont optimisés pour maximiser le critère joint, les séries temporelles sont "interprétées" de manière différente par rapport à une simple régression sur processus gaussien.

paramètres sont optimisés de manière à retranscrire au mieux, la relation imposée par le modèle. En effet, l'inférence va choisir les paramètres qui maximise la probabilité que les données aient été produite selon ce modèle. Les paramètres d'une variable fonctionnelle estimés par l'inférence en 1 étape sont donc potentiellement différents de ceux de la méthode en deux étapes.

Calcul du gradient. On ne réécrira pas le gradient relativement à α et b vu dans la partie **6.6**. Concentrons nous à présent sur le gradient relatif aux paramètres fonctionnels. $\forall i \in [|1, N|]$ et pour tout indice k de l'un des paramètres de la fonction de covariance du processus gaussien F^i nous avons :

$$\begin{aligned}
\frac{\partial \mathcal{L}^p_\lambda(Y, X|\mathcal{P})}{\partial \theta^i_k} &= \frac{\partial \mathcal{L}(Y|X, \mathcal{P})}{\partial \theta^i_k} + \frac{\partial \mathcal{L}(X|\mathcal{P})}{\partial \theta^i_k} \\
&= \frac{1}{\mathbb{V}_i}\left(-2\frac{\partial \mathbb{E}_i}{\partial \theta^i_k}(Y_i - \mathbb{E}_i) + \left(1 - \frac{(Y_i - \mathbb{E}_i)^2}{\mathbb{V}_i}\right)\frac{\partial \mathbb{V}_i}{\partial \theta^i_k}\right) \\
&\quad -(X^i - \eta^i)^T . \Psi^i . \frac{\partial K^i}{\partial \theta^i_k} . \Psi^i . (X^i - \eta^i) + Tr\left(\Psi^i \frac{\partial K^i}{\partial \theta^i_k}\right)
\end{aligned}$$

$$\begin{aligned}
\text{Où } \frac{\partial \mathbb{E}_i}{\partial \theta^i} &= \frac{1}{T_i}\int_0^{T_i} \frac{\partial \mathbb{E}(F^i(s)|X^i)}{\partial \theta^i} \beta_j(s, T_i) ds \\
&= \frac{1}{T_i}\int_0^{T_i} \left(\frac{\partial K^i_*(s)^T}{\partial \theta^i} - K^i_*(s)^T \Psi^i \frac{\partial K^i}{\partial \theta^i}\right) \Psi^i (X^i - \eta^i)\, \beta(s, T_i) ds
\end{aligned}$$

$$\text{et } \frac{\partial \mathbb{V}_i}{\partial \theta^i} = \frac{1}{T_i^2}\int_{[0,T_i]^2} \frac{\partial \mathbb{C}ov(F^i(s),F^i(t)|X^i)}{\partial \theta^i}\beta(s,T_i)\beta(t,T_i)dsdt$$
$$= \frac{1}{T_i^2}\int_{[0,T_i]^2}\left(\frac{\partial k^i(s,t)}{\partial \theta^i} + K_*^{iT}(s)\Psi^i\left(\frac{\partial K^i}{\partial \theta^i}\Psi^i K_*^i(t) - 2\frac{\partial K_*^i(t)}{\partial \theta^i}\right)\right)\beta(s,T_i)\beta(t,T_i)dsdt$$

$$\frac{\partial \mathcal{L}_\lambda^p(Y,X|\mathcal{P})}{\partial \gamma^i} = \frac{1}{\mathbb{V}_i}\left(-2\frac{\partial \mathbb{E}_i}{\partial \gamma^i}(Y_i-\mathbb{E}_i) + \left(1-\frac{(Y_i-\mathbb{E}_i)^2}{\mathbb{V}_i}\right)\frac{\partial \mathbb{V}_i}{\partial \gamma^i}\right)$$
$$-(X^i-\eta^i)^T.\Psi^{i2}.(X^i-\eta^i) + Tr\left(\Psi^i\right)$$

$$\text{Où } \frac{\partial \mathbb{E}_i}{\partial \gamma_j^i} = \frac{1}{T_i}\int_0^{T_i}\frac{\partial \mathbb{E}(F_j^i(s)|X_j^i)}{\partial \gamma_j^i}\beta_j(s,T_i)ds$$
$$= \frac{1}{T_i}\int_0^{T_i} K_*^{ijT}(s).\Psi_j^{i2}.(X_j^i-\eta_j^i)\beta_j(s,T_i)ds$$
$$\text{et } \frac{\partial \mathbb{V}_i}{\partial \gamma_j^i} = \frac{1}{T_i^2}\int_{[0,T_i]^2}\frac{\partial \mathbb{C}ov(F_j^i(s),F_j^i(t)|X_j^i)}{\partial \gamma_j^i}\beta_j(s,T_i)\beta_j(t,T_i)dsdt$$
$$= \frac{1}{T_i^2}\int_{[0,T_i]^2}(k_j^i(s,t) - K_*^{ijT}(s).\Psi_j^{i2}.K_*^{ij}(s))\beta_j(s,T_i)\beta_j(t,T_i)dsdt$$

$$\frac{\partial \mathcal{L}_\lambda^p(Y,X|\mathcal{P})}{\partial \eta_j^i} = -2\frac{\partial \mathbb{E}_i}{\partial \eta^i}\frac{(Y_i-\mathbb{E}_i)}{\mathbb{V}_i} - 2X^{iT}\Psi^i 1_{m_i} + 2\eta^i 1_{m_i}^T\Psi^i 1_{m_i}$$

$$\text{Où } \frac{\partial \mathbb{E}_i}{\partial \eta^i} = \frac{1}{T_i}\int_0^{T_i}\frac{\partial \mathbb{E}(F^i(s)|X^i)}{\partial \gamma^i}\beta(s,T_i)ds = \frac{1}{T_i}\int_0^{T_i}(1+K_*^{iT}(s)\Psi^i 1_{m_i})\beta(s,T_i)ds$$

On obtient $\frac{\partial \mathcal{L}(X|\mathcal{P})}{\partial \eta^i}$ en développant, puis en dérivant le polynome $(X^i-\eta^i 1_{m_i})^T\Psi^i(X^i-\eta^i 1_{m_i}) = X^{iT}\Psi^i X^i - 2\eta^i X^{iT}\Psi^i 1_{m_i} + \eta^{i2}1_{m_i}^T\Psi^i 1_{m_i}$.

Coût d'évaluation du gradient. Comme la densité d'échantillonnage est faible, m_i est supposé petit et l'inversion de Ψ^i n'est pas forcément la partie la plus coûteuse en calcul, sa complexité est $C_{inversion} = O(m_i^3)$. Un autre obstacle obstacle computationnel inflexible pour évaluer le gradient de manière réside dans le calcul de $\mathbb{V}_i$ et de ses dérivées relativement à chaque paramètre fonctionnel, à cause de l'approximation des intégrales double de k^i et ses dérivées partielles. En effet, lorsque Ψ^i est calculée, la complexité $C_{variance}$ de l'approximation de chacune des intégrales double est de l'ordre de J^2. Or, m_i est à priori petit devant J car on considère que la densité d'échantillonnage est faible. On en déduit que $C_{variance}$ est grand devant m_i^2. En prenant par exmeple J de l'ordre de m_i^2, on se retrouve avec $C_{variance}$ de l'ordre de m_i^4 qui est donc grand devant $C_{inversion}$. Précisons que ce calcul de la variance doit être mené N fois, tout comme l'inversion matricielle. Si le temps de calcul dépend de la finesse d'approximation souhaitée, c'est, en pratique, bien l'élément limitant pour l'évaluation du critère et du gradient. Enfin, l'évaluation de β en chaque point de la grille d'intégration est un coût fixe qui peut être élevé si la taille K_β de la base de fonctions qui le représente est grande. Cela dit, ce coût est contourné en évaluant et stockant chaque fonction ϕ_k de la base fonctionnelle sur la grille d'intégration lors d'une étape préliminaire à l'optimisation. Cette méthode représente une dépense de

mémoire substantielle mais abordable pour la finesse souhaitée et l'ordre de grandeur de nos jeux de données.

Algorithme du gradient stochastique. On utilise un algorithme du gradient stochastique, comme dans la partie **6.5**, pour optimiser ce critère coûteux en calcul lorsque N devient grand. Cependant, remarquons que cette fois-ci, il faut choisir un nombre d'itération maximale adapté à la taille de l'échantillon, car les paramètres de VFT d'un individu ne sont mis à jour qu'aux itérations où cet individu est choisi. Par exemple, si on choisit un nombre d'itération inférieur à N, le nombre de mise à jour des paramètres fonctionnels d'un individu est inférieur à 1 en espérance.

6 Étude des performances à partir de données artificielles

On compare les différentes méthodes développées précédemment par une étude sur données simulées en langage Python. Pour rester simple, cette étude empirique considère une seule variable fonctionnelle et aucune variable scalaire. Nous avons voulu tester les performances des méthodes présentées plus haut en fonction de deux facteurs décrits dans la partie **2**, la densité d'échantillonnage et l'erreur de mesure sur la variable fonctionnelle. Nous faisons aussi varier la taille de l'échantillon pour observer la consistance empirique des estimateurs. Tout le code nécessaire à la production de ces résultats est accessible à l'adresse : $https : //github.com/ChrisBotella/GPVDFR$. Le script de l'expérience se trouve dans le fichier `Experience.ipynb` (jupyter notebook), les fonctions python nécessaires à la simulation de données sont contenues dans `Simu.py` et celles liées à l'inférence dans `NLL.py`.

6.1 Procédure de simulation de données

Simulation des temps d'observation des variables fonctionnelles On définit deux paramètres, la densité d'échantillonnage $n \in \mathbb{N}*$, et la taille de l'échantillon $N \in \mathbb{N}*$. On génère les domaines et instants d'observations des N individus selon la procédure suivante.

$\forall i \in [|1, N|]$,

- On tire $T_i \sim \beta(5, 5)$. Ainsi, $T_i \in [0, 1]$.
- Si $nT_i - 2 > 0$, on tire $m \sim \mathcal{P}(nT_i - 2)$, sinon on tire $m \sim P(0.1)$. Puis, on prend $m_i = m + 2$.
- On tire indépendamment $t_1^i, ..., t_{m_i}^i \sim U(0, T_i)$.

Ainsi, on obtient, d'une part, pour chaque variable fonctionnelle une densité d'échantillonnage égale en espérance à n, quel que soit le domaine (tant qu'on ne choisit pas n trop petit). D'autre part, on s'assure d'avoir au moins deux points par série simulée. L'exemple donné en figure 2 est une simulation avec une densité d'échantillonnage de 10 points par unité de temps. On trouve bien une moyenne du nombre de mesures à 5 points, car la distribution des durées de mesure est centrée sur 0.5.

Simulation des mesures de variables fonctionnelles Pour la simulation des valeurs prises par la variable fonctionnelle, nous avons réutilisé le schéma proposé dans Gellar *et al.* [2014], qui est également employé dans Goldsmith *et al.* [2012]. On définit l'erreur-type de mesure $\gamma\mathbb{R}^+$ et on mène à bien la procédure suivante :

$\forall i \in [|1, N|]$,

- On tire $u^i \sim \mathcal{N}(0, 1)$, et $\forall k \in [|1, 10|]$, $v_{1k}^i, v_{2k}^i \sim \mathcal{N}(0, 4/k^2)$. Cela définit le signal de la variable fonctionnelle i par la fonction $s \to F^i(s) = u^i + \sum_{k=0}^{10} v_{1k}^i \, sin\,(2\pi ks) + v_{2k}^i \, cos\,(2\pi ks)$.
- On tire un bruit gaussien i.i.d. pour chaque instant d'observation $\forall j \in [|1, m_i|], \delta_j \sim \mathcal{N}(0, \gamma^2)$ et on prend $X^i(t_j^i) = F^i(t_j^i) + \delta_j$.

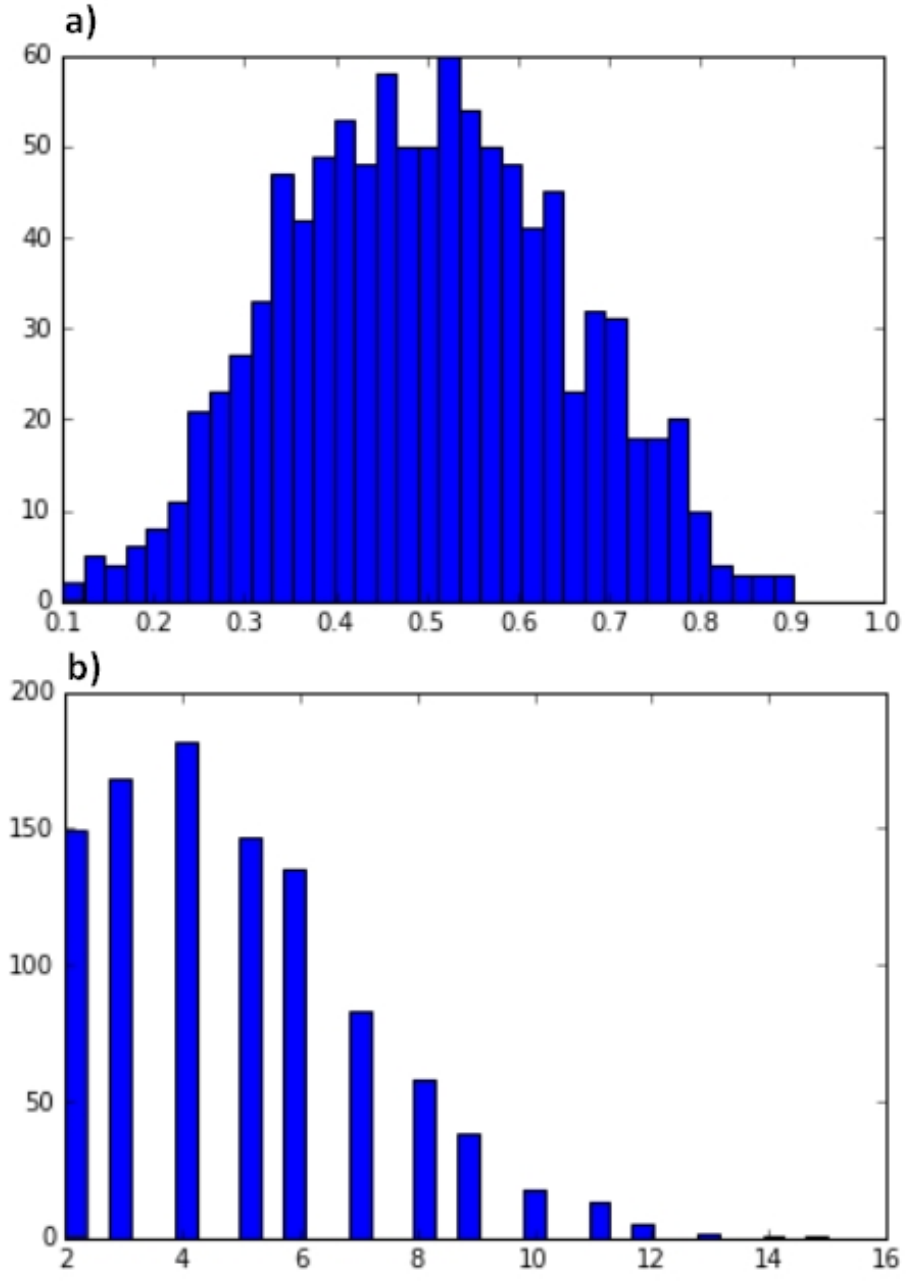

FIGURE 2 – Simulation d'instants et domaines de mesure avec $n = 10$ et $N = 1000$.**a)** distribution des durées de mesure T_i.**b)** distribution des nombres de mesures m_i.

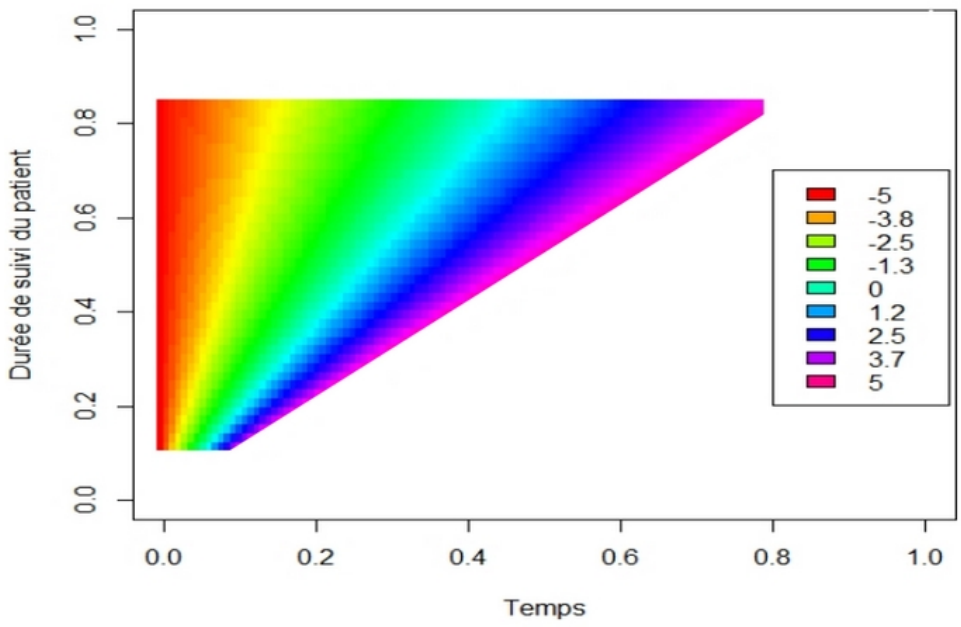

FIGURE 3 – Représentation du paramètre fonctionnel générateur pour le calcul du label sur le domaine $\{0 \leq T \leq 1, 0 \leq t \leq T\}$.

Simulation des labels On utilise le paramètre fonctionnel :

$$t, T \to \beta(s, T) = \frac{10t}{T - 5}$$

Une représentation graphique sur la surface $\{t, T \in [0, 1]^2, t \leq T\}$ est fournie en figure 3. Les labels sont simulés en calculant $\forall i \in [|1, N|]$,

$$Y_i = \int_0^{T_i} \beta(s, T_i) F^i(s) ds$$

L'intégrale est approximée avec la fonction python *scipy.integrate.nquad* avec une précision fixée à 5.10^{-4}.

6.2 Design expérimental

Paramètres testés N est à valeur dans $\{100, 200, 500\}$. n prend les valeurs 10 et 100 ce qui correspond respectivement à un régime de faible et de grande densité d'échantillonnage relativement au potentiel de fluctuation du signal. $\gamma \in \{0, 1\}$ représentant respectivement un régime d'absence ou présence d'erreur de mesure sur la variable fonctionnelle. Ces paramètres aboutissent à 12 combinaisons, et nous simulons 40 jeux de données, appelés répétitions, pour chacune de ces combinaisons.

Évaluation des estimateurs Le critère **ISE** (Integrated Square Error) est l'intégrale du carré de l'erreur d'estimation du paramètre fonctionnel β sur le domaine maximal d'observation, définit pour un estimateur $\hat{\beta}$ par la formule :

$$ISE(\hat{\beta}) = \int_{T_{min}}^{T_{max}} \int_0^t (\beta(s, t) - \hat{\beta}(s, t))^2 ds dt$$

Nous calculons ce critère sur chaque répétition de chaque expérience et chaque estimateur, l'approximant par une moyenne sur grille discrète. Ensuite, nous calculons le critère **MISE** (Mean Integrated Square Error), c'est à dire la moyenne des **ISE**s, sur les répétitions de chaque expérience pour chaque estimateur. Ceci constitue notre principal critère de quantification de performance d'estimation. Nous calculons aussi l'écart-type des **ISE**s pour mesurer empiriquement la variance des estimateurs.

Évaluation des prédictions Nous calculons à chaque inférence réalisée la **rMSE** (root Mean Square Error) sur un jeu de données test indépendant dont les paramètres de simulation sont identiques au jeu de données d'apprentissage. Ainsi, nous aurons une idée de la part de variabilité de Y expliquée par le modèle, et un critère de comparaison entre de la performance sur données simulées et données réelles plus tard.

6.3 Détails supplémentaires sur les méthodes dans le cadre de l'expérience

Noyau choisi pour le modèle VFT. Nous avons choisi pour la modélisation de la fonction de covariance de la variable fonctionnelle dans l'ensemble des méthode une somme de deux noyaux : le premier est un RBF (exponentiel quadratique) et le deuxième est périodique. Ces fonctions sont décrites dans l'appendice. Le premier nous permet de capter une corrélation temporelle générique due à la continuité générale du signal, et le deuxième sert à capter la périodicité principale de ce signal sinusoïdal. Nous connaissons la forme de la fonction génératrice, et celle-ci est une somme de 20 fonctions sinusoïdales à intensité décroissante. Cependant, il n'est pas raisonnable d'utiliser une somme de nombreux noyaux périodiques ici car trop peu de points sont observés pour en inférer correctement les paramètres.

Paramètres de l'inférence. Nous utilisons pour tous les modèles testés une finesse de maille de $J = 70$, et une base fonctionnelle de taille $K_\beta = 25$, ce qui correspond aux combinaisons de monômes de degrés 1 à 5 dans chaque dimension.

Régularisation. En ce qui concerne la valeur de l'hyperparamètre de régularisation λ, nous réaliser une procédure de bootstrap. Nous choisissons à priori un ensemble de valeurs de λ, et, pour chacune d'entre elle, nous choisissons aléatoirement un sous-échantillon du jeu de données comprenant 70% des individus. On réalise alors une inférence des paramètres sur ce sous-échantillon et on calcule la rMSE de prédiction sur les 30% non utilisés pour l'inférence. On retient alors le paramètre estimé associé à la valeur de λ dont la rMSE est la plus faible.

6.4 Résultats comparés

Nous évaluons les 4 méthodes présentées plus haut sur les données simulées, à savoir la méthode des moindres carrés simple, la méthode des moindres carrés pénalisée, la méthode du maximum de vraisemblance en deux étapes et la méthode jointe pénalisée. Le tableau de la figure 4 fournit les résultats obtenus pour chaque combinaison de paramètres de simulation N et n. On y fait figurer le critère **MISE**, son écart-type sur les 40 jeux de données générés, et la moyenne de la **rMSE**. Nous fournissons ces résultats en l'état, même si plusieurs jeux de données simulés n'ont pas pu être testés dans les temps.

La méthode des moindres carré simple procure des résultats abbérants, ce qui est normal au vu de la taille assez faible des jeux de données par rapport aux nombre de paramètres estimés, qui favorise le sur-ajustement. La méthode MC pénalisée donne des résultats meilleurs et plus stables, ce qui montre bien l'intérêt de la régularisation. Les autres méthodes ont des performances du même ordre. Cela dit, la variance du MISE est plus grande pour les méthodes du maximum de vraisemblance. On observe une performance légèrement plus élevé et plus stable lorsque la taille du jeu de données augmente.

		Sans erreur		
	Inference	MISE	std.	rMSE
N=100 et n=10	MC	4.5e3	3.6e3	3.27
	MC pénalisé	8.42	0.53	0.85
	$\mathcal{L}^p(Y\|X)$	8.19	1.33	0.83
	$\mathcal{L}^p(Y,X)$	8.75	1.89	0.82
N=200 et n=10	MC	150.	80.12	1.59
	MC pénalisé	8.33	0.38	0.78
	$\mathcal{L}^p(Y\|X)$	7.83	1.08	0.76
	$\mathcal{L}^p(Y,X)$	8.19	1.21	0.76

FIGURE 4 – Premiers résultats expérimentaux des différentes méthodes sur données simulées.

Problème de convergence des méthodes de vraisemblance En mettant en parallèle ces résultats avec les valeurs finales des gradient globaux, on constate que fréquemment les résultats des méthodes de vraisemblance sont affectés par des problèmes de convergence de l'algorithme stochastique d'optimisation. En effet, l'algorithme du gradient stochastique a pour critère le passage du gradient de l'individu tiré sous un seuil de norme. Cependant, cela ne veut pas toujours dire que le gradient total a été annulé, en particulier lorsque la taille de l'échantillon est grande. Ainsi l'algorithme devrait être amélioré, en finissant par exemple la convergence grâce à une méthode de gradient déterministe. Outre ces problèmes de convergence, les résultats des méthodes de vraisemblance semblent assez proches entre-elles, donc la méthode en deux étape semble globalement plus intéressante, car bien plus rapide et plus facile à optimiser.

Meilleur choix de la base de β Il est fort à parier que ces performances pourraient être améliorer en chosissant une base fonctionnelle plus adaptée à la régression fonctionnelle en domaine variable. Gellar *et al.* [2014] évoque les bases de thin plate splines [Wood, 2003], mais on pourrait aussi envisager d'adapter des bases de B-splines en deux dimensions. Les fonctions de ces bases ont l'avantage notoire d'avoir des support qui se recoupent peu, assurant par exemple la définie positivité de la matrice de pénalisation I_p et des matrices de variance V^i (et donc la coercivité du critère), mais cela procurer bien d'autres avantages.

7 Modélisation de la censure

7.1 Introduction de la vraisemblance du modèle censuré

A présent, on revient au modèle théorique et on introduit la censure des données dont le mécanisme a été présenté dans la partie **1**. Rappelons les notations correspondantes, $\forall i \in [|1, N|]$:

- C_i est la variable réelle du temps de censure du patient i, soit la durée entre le moment où le patient quitte l'hôpital (fin de la période de mesure) et le moment où il quitte la cohorte (l'instant à partir duquel on n'a plus aucune information sur son devenir).
- $\tau_i^c = \min(\tau_i, C_i)$ est le temps censuré de ré-admission hospitalière.
- δ_i est l'indicateur de non-censure, $\delta_i = 1$ si $\tau_i^c = \tau_i$ ou O sinon.
- $\tilde{C}_i := g(C_i)$.
- $Y_i^c := g(\tau_i^c) = \min\left(g(\tau_i), g(C_i)\right) = \min(Y_i, \tilde{C}_i)$ le label censuré.

On considère alors de nouvelles hypothèses sur ce modèle censuré.

Hypothèses :

1. $\forall i \in [|1, N|], \tilde{C}_i \perp\!\!\!\perp Y_i \mid \left(X^i, Z_i\right)$ et $\tilde{C}_i \perp\!\!\!\perp \left(X^i, Z_i\right)$.
2. $\tilde{C}_i$ admet une densité g. On note G sa fonction de répartition définie sur $\mathbb{R}$ et $\bar{G} = 1 - G$.

L'hypothèse 8. est classique en analyse de survie [Klein & Moeschberger, 2005]. On s'intéresse alors au modèle **(4)**, qui complète le modèle **(3)** et l'ensemble de ses hypothèses par l'ajout des définitions et hypothèses ci-dessus liées à la censure. $\forall i \in [|1, N|]$:

$$\textbf{(4)} \begin{cases} Y_i^c = \min(Y_i, \tilde{C}_i) \\ Y_i = Z_i^T \alpha + \dfrac{1}{T_i} \displaystyle\sum_{j=1}^{D} \int_0^{T_i} F_j^i(s) \beta_j(s, T_i) ds + \epsilon_i \\ X_j^i(t_j^i) = F_j^i(t_j^i) + \zeta_j^i \ , \ \forall j \in [|1, D|] \end{cases}$$

On rappelle, comme on l'a vu au **chapitre 5**, que Y_i admet une densité conditionnellement aux variables fonctionnelles $X_1^i, ..., X_D^i$ qui est connue et explicite. On note $F(.|X^i, \mathcal{P})$ sa fonction de répartition définie sur $\mathbb{R}$ et $\bar{F}(.|X^i) = 1 - F(.|X^i, \mathcal{P})$.

A présent, comme l'hypothèse 6. tient toujours, les observations sont indépendantes et l'on pourra toujours décomposer la fonction de vraisemblance globale des données en un produit des fonctions de vraisemblance de chaque observation. Exprimons $\forall i \in [|1, N|]$ la vraisemblance jointe censurée du patient i en la décomposant en un terme de vraisemblance de variables fonctionnelles, que l'on connaît bien car on l'a exprimé dans le **chapitre 5**, et un terme de vraisemblance conditionnelle jointe censurée : $p(Y_i^c, \delta_i, X^i|\mathcal{P}) = p(Y_i^c, \delta_i|X^i, \mathcal{P}) p(X^i|\mathcal{P})$. Il ne nous reste plus qu'à calculer $p(Y_i^c, \delta_i|X^i, \mathcal{P})$ à partir de $p(Y_i|X^i, \mathcal{P})$.

Vraisemblance jointe conditionnelle censurée. Commençons par écrire la vraisemblance locale jointe de $(Y_i^c, \delta_i), \forall i \in [|1, N|]$. Si $\delta_i = 1$, comme $\delta_i = 1_{\{Y_i \leq \tilde{C}_i\}}$, on a $Y_i^c = Y_i \leq \tilde{C}_i$. Comme $Y_i \perp\!\!\!\perp \tilde{C}_i \mid X^i$, on peut factoriser la vraisemblance d'une observation $y_i^c \in \mathbb{R}$ de la variable Y_i (on distingue ici la variable aléatoire et une observation pour les besoins de clarté du calcul) en utilisant les densités de $Y_i|X^i$ et $\tilde{C}_i$ de la manière suivante :

$$p(Y_i^c = y_i^c, \delta_i = 1|X^i, \mathcal{P}) = p(Y_i = y_i^c, \tilde{C}_i > y_i^c|X^i, \mathcal{P}) = p(Y_i = y_i^c|X^i, \mathcal{P}, \tilde{C}_i > y_i^c)\ p(\tilde{C}_i > y_i^c)$$

$$= p(Y_i = y_i^c|X^i, \mathcal{P})\, \bar{G}(y_i^c)$$

En suivant un raisonnement similaire, on obtient $p(Y_i^c = y_i^c, \delta_i = 0) = g(y_i^c)\bar{F}(y_i^c|X^i, \mathcal{P})$. Cela nous permet d'écrire la vraisemblance conditionnelle locale d'une observation i (On confond de nouveau l'écriture de la variable aléatoire et de son observation) :

$$p(Y_i^c, \delta_i|X^i, \mathcal{P}) = p(Y_i^c, \delta_i = 1)^{\delta_i} p(Y_i^c, \delta_i = 0)^{1-\delta_i}$$

$$\Leftrightarrow p(Y_i^c, \delta_i|X^i, \mathcal{P}) = p(Y_i^c|X^i, \mathcal{P})^{\delta_i} \bar{F}(Y_i^c|X^i, \mathcal{P})^{1-\delta_i}\ g(Y_i^c)^{1-\delta_i} \bar{G}(Y_i^c)^{\delta_i}$$

Ainsi,

$$p(Y^c, \delta|X, \mathcal{P}) = \prod_{i=1}^{N} p(Y_i^c|X^i, \mathcal{P})^{\delta_i} \bar{F}(Y_i^c|X^i, \mathcal{P})^{1-\delta_i} \prod_{i=1}^{N} g(Y_i^c)^{1-\delta_i} \bar{G}(Y_i^c)^{\delta_i}$$

Log-Vraisemblance négative simplifiée. Classiquement, on choisit de passer au logarithme pour se ramener à une expression de somme :

$$\log(p(Y^c, \delta | X, \mathcal{P})) =$$

$$\sum_{i=1}^{N} \delta_i \log(p(Y_i^c | X^i, \mathcal{P})) + (1-\delta_i) \log(\bar{F}(Y_i^c | X^i, \mathcal{P})) \; + \; \sum_{i=1}^{N} (1-\delta_i) \log(g(Y_i^c)) + \delta_i \log(\bar{G}(Y_i^c))$$

On constate que lorsque la densité de g est fixée, comme on ne s'y intéresse pas et qu'elle est indépendante des paramètres $\mathcal{P}$ que l'on souhaite estimer, maximiser la log-Vraisemblance relativement à $\mathcal{P}$ est indépendant de la deuxième somme, dont on peut donc se passer.
On voudra alors minimiser la log-vraisemblance négative délestée de ce terme :

$$-\log(p(Y^c, \delta | X, \mathcal{P})) = -\sum_{i=1}^{N} \delta_i \log(p(Y_i^c | X^i, \mathcal{P})) + (1-\delta_i) \log(\bar{F}(Y_i^c | X^i, \mathcal{P}))$$

Alors, on remplace par les expressions obtenues dans **le chapitre 5**, on met de côté, comme précédemment, les termes indépendants de $\mathcal{P}$ et on multiplie par 2. On obtient le critère d'intérêt pour l'optimisation des paramètres du modèle **(4)** :

$$\mathcal{L}(Y^c, \delta, X | \mathcal{P}) = \mathcal{L}(Y^c, \delta | X, \mathcal{P}) + \mathcal{L}(X | \mathcal{P}) =$$

$$\sum_{i=1}^{N} \left[\delta_i \left(\frac{(Y_i^c - \mathbb{E}(Y_i | X^i))^2}{\mathbb{V}(Y_i | X^i)} + \log \mathbb{V}(Y_i | X^i) \right) - 2(1-\delta_i) \; \log(\bar{F}(Y_i^c | X^i, \mathcal{P})) \right.$$
$$\left. + \sum_{j=1}^{D} \left(X_j^i(t_j^i) - \eta_j^i\right)^T . \Psi_j^i . \left(X_j^i(t_j^i) - \eta_j^i\right) + \log(\det(\Psi_j^{i-1})) \right]$$

7.2 Calcul du gradient de la vraisemblance

Soit une variable aléatoire réelle telle que $G \sim \mathcal{N}(0,1)$.
On utilisera dans les calculs de gradients qui suivent la propriété suivante sur la dérivation du logarithme d'une fonction de répartition gaussienne.

Propriété :
Soit $Y \sim \mathcal{N}(\mu, \sigma^2)$ où μ et σ^2 sont des fonction d'une variable $x \in \mathbb{R}$. Alors la dérivée du logarithme de la fonction de répartition de Y par rapport à x peut s'exprimer comme un polynôme de second ordre des moments conditionnels d'une variable normale centrée réduite. $\forall a \in \mathbb{R}$ on a :

$$\frac{\partial \log \mathbb{P}(Y \leq a)}{\partial x} = \frac{1}{2} \mathbb{E}\left(G^2 \mid G \leq \frac{a-\mu}{\sigma}\right) \frac{1}{\sigma^2} \frac{\partial \sigma^2}{\partial x} + \mathbb{E}\left(G \mid G \leq \frac{a-\mu}{\sigma}\right) \frac{1}{\sigma} \frac{\partial \mu}{\partial x} - \frac{1}{2} \frac{1}{\sigma^2} \frac{\partial \sigma^2}{\partial x}$$

Preuve :
Il suffit de calculer la dérivée de la densité gaussienne continue dans l'intégrale et de remarquer l'apparition des moments gaussiens.

$$\frac{\partial \log \mathbb{P}(Y \leq a)}{\partial x} = \frac{1}{\mathbb{P}(Y \leq a)} \int_{-\infty}^{a} \frac{\partial}{\partial x} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu)^2}{2\sigma^2}} \right) dy$$
$$= \frac{1}{\mathbb{P}(Y \leq a)} \int_{-\infty}^{a} \left[\frac{1}{2} \frac{(y-\mu)^2}{\sigma^4} \frac{\partial \sigma^2}{\partial x} + \frac{(y-\mu)}{\sigma^2} \frac{\partial \mu}{\partial x} - \frac{1}{2} \frac{1}{\sigma^2} \frac{\partial \sigma^2}{\partial x} \right] \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy$$

$$= \frac{1}{\mathbb{P}(G \leq \frac{a-\mu}{\sigma})} \int_{-\infty}^{\frac{a-\mu}{\sigma}} \left[\frac{1}{2} \frac{U^2}{\sigma^2} \frac{\partial \sigma^2}{\partial x} + \frac{U}{\sigma} \frac{\partial \mu}{\partial x} - \frac{1}{2} \frac{1}{\sigma^2} \frac{\partial \sigma^2}{\partial x} \right] \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-U^2}{2}} \sigma dU$$

Avec le changement de variable $U = \frac{y-\mu}{\sigma}$. On obtient alors le résultat en faisant apparaître la densité conditionnelle grâce à la formule des probabilités conditionnelles :

$$= \int_{-\infty}^{\frac{a-\mu}{\sigma}} \left[\frac{1}{2} \frac{U^2}{\sigma^2} \frac{\partial \sigma^2}{\partial x} + \frac{U}{\sigma} \frac{\partial \mu}{\partial x} - \frac{1}{2} \frac{1}{\sigma^2} \frac{\partial \sigma^2}{\partial x} \right] \frac{f_G(U)}{\mathbb{P}(G \leq \frac{a-\mu}{\sigma})} dU$$

$$= \frac{1}{2} \mathbb{E}\left(G^2 | G \leq \frac{a-\mu}{\sigma} \right) \frac{1}{\sigma^2} \frac{\partial \sigma^2}{\partial x} + \mathbb{E}\left(G | G \leq \frac{a-\mu}{\sigma} \right) \frac{1}{\sigma} \frac{\partial \mu}{\partial x} - \frac{1}{2} \frac{1}{\sigma^2} \frac{\partial \sigma^2}{\partial x} \quad \square$$

On note dans tous les calculs qui suivent et $\forall i \in [|1, N|]$, $\mathbb{E}_i := \mathbb{E}(Y_i | X^i, \mathcal{P})$, $\mathbb{V}_i := \mathbb{V}(Y_i | X^i, \mathcal{P})$ et $\tilde{Y_i^c} = \frac{Y_i^c - \mathbb{E}_i}{\sqrt{\mathbb{V}_i}}$. On dérive les gradients de $\mathcal{L}_\lambda^p(Y, X | \mathcal{P})$ relativement aux différents paramètres du modèle **(4)** à l'aide de cette propriété.

- $\frac{\partial \mathcal{L}_\lambda^p(Y, X|\mathcal{P})}{\partial \alpha} = \frac{\partial \mathcal{L}(Y|X, \mathcal{P})}{\partial \alpha} =$

$$\sum_{i=1}^{N} -2 \left(\delta_i \frac{\tilde{Y_i^c}}{\sqrt{\mathbb{V}_i}} + (1 - \delta_i) \frac{\mathbb{E}(G \mid G \geq \tilde{Y_i^c})}{\sqrt{\mathbb{V}_i}} \right) Z_i$$

- $\forall j \in [|1, D|]$, $\frac{\partial \mathcal{L}_\lambda^p(Y, X|\mathcal{P})}{\partial b_j} = \frac{\partial \mathcal{L}(Y|X, \mathcal{P})}{\partial b_j} + \lambda \frac{\partial Pen(\beta)}{\partial b_j}$

Où $\frac{\partial Pen(\beta)}{\partial b_j} = 2(I_s + 2I_c + I_t) b_j$

et $\frac{\partial \mathcal{L}(Y|X, \mathcal{P})}{\partial b_j} =$

$$\sum_{i=1}^{N} \left[\delta_i \left(-2 \frac{\partial \mathbb{E}_i}{\partial b_j} \frac{\tilde{Y_i^c}}{\sqrt{\mathbb{V}_i}} + \left(\frac{1 - \tilde{Y_i^{c2}}}{\mathbb{V}_i} \right) \frac{\partial \mathbb{V}_i}{\partial b_j} \right) - \right.$$

$$\left. 2(1 - \delta_i) \left(\frac{1}{2} \mathbb{E}(G^2 | G \geq \tilde{Y_i^c}) \frac{1}{\mathbb{V}_i} \frac{\partial \mathbb{V}_i}{\partial b_j} + \mathbb{E}(G | G \geq \tilde{Y_i^c}) \frac{1}{\sqrt{\mathbb{V}_i}} \frac{\partial \mathbb{E}_i}{\partial b_j} - \frac{1}{2} \frac{1}{\mathbb{V}_i} \frac{\partial \mathbb{V}_i}{\partial b_j} \right) \right]$$

Avec $\frac{\partial \mathbb{E}_i}{\partial b_j} = \frac{1}{T_i} \int_0^{T_i} \mathbb{E}(F_j^i(s) | X_j^i) \phi(s, T_i) ds = \frac{1}{T_i} \int_0^{T_i} \phi(s, T_i) \left(\eta_j^i + K_*^{ijT}(s) . \Psi_j^i . (X_j^i - \eta_j^i) \right) ds$

et $\frac{\partial \mathbb{V}_i}{\partial b_j} = \frac{1}{T_i^2} \int_{[0,T_i]^2} \mathbb{C}ov \left(F_j^i(s), F_j^i(t) | X_j^i \right) \left(\phi(s, T_i) \beta_j(t, T_i) + \phi(t, T_i) \beta_j(s, T_i) \right) ds dt$

$= \frac{2}{T_i^2} \int_{[0,T_i]^2} \left(k_j^i(s,t) - K_*^{ijT}(s) . \Psi_j^i . K_*^{ij}(t) \right) \phi(s, T_i) \beta_j(t, T_i) \, ds dt$

- $\forall j \in [|1, D|], \forall i \in [|1, N|]$, $\frac{\partial \mathcal{L}_\lambda^p(Y, X|\mathcal{P})}{\partial \theta_j^i} = \frac{\partial \mathcal{L}(Y|X, \mathcal{P})}{\partial \theta_j^i} + \frac{\partial \mathcal{L}(X|\mathcal{P})}{\partial \theta_j^i} =$

$$\frac{\delta_i}{\mathbb{V}_i} \left(-2 \frac{\partial \mathbb{E}_i}{\partial \theta_j^i} (Y_i^c - \mathbb{E}_i) + \left(1 - \frac{(Y_i^c - \mathbb{E}_i)^2}{\mathbb{V}_i} \right) \frac{\partial \mathbb{V}_i}{\partial \theta_j^i} \right)$$

$$-2(1 - \delta_i) \left(\frac{1}{2} \mathbb{E}(G^2 | G \geq \tilde{Y_i^c}) \frac{1}{\mathbb{V}_i} \frac{\partial \mathbb{V}_i}{\partial \theta_j^i} + \mathbb{E}(G | G \geq \tilde{Y_i^c}) \frac{1}{\sqrt{\mathbb{V}_i}} \frac{\partial \mathbb{E}_i}{\partial \theta_j^i} - \frac{1}{2} \frac{1}{\mathbb{V}_i} \frac{\partial \mathbb{V}_i}{\partial \theta_j^i} \right)$$

$$-(X_j^i - \eta_j^i)^T . \Psi_j^i . \frac{\partial K_j^i}{\partial \theta_j^i} . \Psi_j^i . (X_j^i - \eta_j^i) + Tr \left(\Psi_j^i \frac{\partial K_j^i}{\partial \theta_j^i} \right)$$

Où $\frac{\partial \mathbb{E}_i}{\partial \theta_j^i} = \frac{1}{T_i}\int_0^{T_i} \frac{\partial \mathbb{E}(F_j^i(s)|X_j^i)}{\partial \theta_j^i}\beta_j(s,T_i)ds$

$= \frac{1}{T_i}\int_0^{T_i}\left(\frac{\partial K_*^{ij}(s)^T}{\partial \theta_j^i} - K_*^{ij}(s)^T\Psi_j^i\frac{\partial K_j^i}{\partial \theta_j^i}\right)\Psi_j^i(X_j^i-\eta_j^i)\,\beta_j(s,T_i)ds$

et $\frac{\partial \mathbb{V}_i}{\partial \theta_j^i} = \frac{1}{T_i^2}\int_{[0,T_i]^2}\frac{\partial \mathbb{C}ov(F_j^i(s),F_j^i(t)|X_j^i)}{\partial \theta_j^i}\beta_j(s,T_i)\beta_j(t,T_i)dsdt$

$= \frac{1}{T_i^2}\int_{[0,T_i]^2}\left(\frac{\partial k_j^i(s,t)}{\partial \theta_j^i} - 2K_*^{ijT}(s)\,\Psi_j^i\,\frac{\partial K_*^{ij}(t)}{\partial \theta_j^i} + K_*^{ijT}(s)\,\Psi_j^i\,\frac{\partial K_j^i}{\partial \theta_j^i}\,\Psi_j^i\,K_*^{ij}(t)\right)\beta_j(s,T_i)\beta_j(t,T_i)dsdt$

- $\frac{\partial \mathcal{L}_\lambda^p(Y,X|\mathcal{P})}{\partial \gamma_j^i} =$

$$\frac{\delta_i}{\mathbb{V}_i}\left(-2\frac{\partial \mathbb{E}_i}{\partial \gamma_j^i}(Y_i^c-\mathbb{E}_i) + \left(1-\frac{(Y_i^c-\mathbb{E}_i)^2}{\mathbb{V}_i}\right)\frac{\partial \mathbb{V}_i}{\partial \gamma_j^i}\right)$$

$$-2(1-\delta_i)\left(\frac{1}{2}\,\mathbb{E}(G^2|G\geq \tilde{Y_i^c})\,\frac{1}{\mathbb{V}_i}\frac{\partial \mathbb{V}_i}{\partial \gamma_j^i} + \mathbb{E}(G|G\geq \tilde{Y_i^c})\,\frac{1}{\sqrt{\mathbb{V}_i}}\frac{\partial \mathbb{E}_i}{\partial \gamma_j^i} - \frac{1}{2}\,\frac{1}{\mathbb{V}_i}\frac{\partial \mathbb{V}_i}{\partial \gamma_j^i}\right)$$

$$-(X_j^i-\eta_j^i)^T.\Psi_j^{i2}.(X_j^i-\eta_j^i) + Tr\left(\Psi_j^i\right)$$

Où $\frac{\partial \mathbb{E}_i}{\partial \gamma_j^i} = \frac{1}{T_i}\int_0^{T_i}\frac{\partial \mathbb{E}(F_j^i(s)|X_j^i)}{\partial \gamma_j^i}\beta_j(s,T_i)ds = \frac{1}{T_i}\int_0^{T_i}K_*^{ijT}(s).\Psi_j^{i2}.(X_j^i-\eta_j^i)\beta_j(s,T_i)ds$

et $\frac{\partial \mathbb{V}_i}{\partial \gamma_j^i} = \frac{1}{T_i^2}\int_{[0,T_i]^2}\frac{\partial \mathbb{C}ov(F_j^i(s),F_j^i(t)|X_j^i)}{\partial \gamma_j^i}\beta_j(s,T_i)\beta_j(t,T_i)dsdt$

$= \frac{1}{T_i^2}\int_{[0,T_i]^2}(k_j^i(s,t) - K_*^{ijT}(s).\Psi_j^{i2}.K_*^{ij}(s))\beta_j(s,T_i)\beta_j(t,T_i)dsdt$

- $\frac{\partial \mathcal{L}_\lambda^p(Y,X|\mathcal{P})}{\partial \eta_j^i} =$

$$-2\delta_i\frac{\partial \mathbb{E}_i}{\partial \eta_j^i}\frac{(Y_i^c-\mathbb{E}_i)}{\mathbb{V}_i} - 2(1-\delta_i)\frac{1}{\sqrt{\mathbb{V}_i}}\frac{\partial \mathbb{E}_i}{\partial \eta_j^i}\mathbb{E}\left(G\,|\,G\geq\tilde{Y_i^c}\right) - 2X_j^{iT}\Psi_j^i 1_{m_ij} + 2\eta_j^i 1_{m_ij}^T\Psi_j^i 1_{m_ij}$$

Où $\frac{\partial \mathbb{E}_i}{\partial \eta_j^i} = \frac{1}{T_i}\int_0^{T_i}\frac{\partial \mathbb{E}(F_j^i(s)|X_j^i)}{\partial \gamma_j^i}\beta_j(s,T_i)ds = \frac{1}{T_i}\int_0^{T_i}(1+K_*^{ijT}(s)\Psi_j^i 1_{m_{ij}})\beta_j(s,T_i)ds$

On obtient $\frac{\partial \mathcal{L}(X|\mathcal{P})}{\partial \eta_j^i}$ en développant, puis en dérivant le polynome $(X_j^i-\eta_j^i 1_{m_ij})^T\Psi_j^i(X_j^i-\eta_j^i 1_{m_ij}) = X_j^{iT}\Psi_j^i X_j^i - 2\eta_j^i X_j^{iT}\Psi_j^i 1_{m_ij} + \eta_j^{i2}1_{m_ij}^T\Psi_j^i 1_{m_ij}$.

8 Conclusion

Ce travail de stage en statistiques appliquées a permis, à partir de l'exemple des problèmes de ré-hospitalisation, de mettre en exergue certaines spécificités des données médicales longitudinales, rarement prises en compte dans les modèles statistique à l'heure actuelle. Cela a donné lieu à la recherche d'une modélisation statistique adaptée. Le panel d'outils offerts dans la littérature sur la régression fonctionnelle a été un terreau fertile, auquel nous avons intégré la modélisation de variables fonctionnelles par des processus gaussiens qui sont un outil populaire aujourd'hui en modélisation

de séries temporelles, en statistiques spatiales et au-delà. En effet, nous avons vu qu'ils permettent de s'adapter aux mieux à des données de faible résolution en modélisant la covariance temporelle avec un nombre de paramètres restreint. Quelques propriétés théoriques de ce modèle de régression linéaire fonctionnelle avec prior gaussien ont été étudiées, et plusieurs méthodes d'inférence proposées et implémentées sous python.
Nous avons alors évalué les performances de ces méthodes sur des données simulées. La méthode des moindres carrés simple se confronte au problème de variance et c'est aussi le cas pour les autres modèles lorsqu'ils ne sont pas pénalisés, ce qui prouve l'intérêt de la régularisation choisie. Les méthodes d'inférence pénalisées ne montrent pas de différence significative de performance entre-elles. Cependant, de grandes améliorations de performance pourraient encore être réalisées, par exemple sur la partie modélisation en choisissant une base fonctionnelle plus adaptée comme les B-splines, courantes en régression fonctionnelle, et sur la partie optimisation en améliorant l'algorithme stochastique qui ne permet pas de convergence totale vers un optimum global en un temps raisonnable pour ce modèle. Enfin, nous avons étendu le cadre de ce modèle à des données censurées pour le raccorder à la problématique initiale. Du travail reste à faire durant la suite de ce stage afin d'appliquer le modèle censuré à des données hospitalières. Au niveau théorique, une question intéressante à adresser serait l'incertitude des estimateurs des paramètres fonctionnels en fonction du nombre d'individus de la base de données, mais à densité d'échantillonnage des VFT constante.

9 Remerciements

Je remercie mes deux maîtres de stage, Agathe Guilloux et Anne-Sophie Jannot, ainsi que Simon Bussy, doctorant à l'UPMC qui ont suivi et guidé mon stage. Celui-ci fut une expérience fort enrichissante.

10 Appendice

10.1 Régression sur processus gaussien

Introduction du modèle On s'intéresse à des observations d'un paramètre vital X à valeurs réelles observé au cours du temps chez un patient. Le nombre et la fréquence des observations au cours du temps est quelconque. Supposons qu'il est mesuré à n instants $t = (t_1....t_n) \in \mathbb{R}+^n$ où il prend les valeurs $X = (X(t_1)...X(t_n)) \in \mathbb{R}^n$. On va modéliser X comme une fonction continue du temps à laquelle s'ajoute un bruit indépendant. Pour cela, nous avons besoin de choisir un ensemble de fonctions admissibles parmi lesquelles on pourra choisir la plus "pertinente" au vu de nos données. L'approche présentée dans l'article Pimentel *et al.* [2013], qui a inspiré le travail de stage est une méthode appelée régression sur processus gaussien. Elle consiste à supposer que la fonction X est générée par une trajectoire d'un processus gaussien altérée par un bruit gaussien indépendant. On modélise alors X de la manière suivante, pour tout $i \in \{1...n\}$:

$$X(t_i) = F(t_i) + \epsilon_i$$

$$\text{où } \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2) \text{ et } F \sim \mathcal{GP}(m, k)$$

les $(\epsilon_i)_{i\in\{1...n\}}$ sont iid. m est la fonction moyenne $\mathbb{R}^+ \to \mathbb{R}$ du processus gaussien (GP) que l'on suppose constante ici. On prend $m = 0$ sans perte de généralité. Le paramètre $k : \mathbb{R}+^2 \to \mathbb{R}+$ est la fonction de covariance de ce processus.

Rappel sur les processus gaussiens et définition de la covariance Les processus gaussiens sont un cas particulier de processus stochastiques homogènes en temps étudiés en théorie des probabilités depuis presque un siècle, voir par exemple [Doob, 1944], et dont les applications en statistiques et en apprentissage automatique sont aujourd'hui communes. Une synthèse réputée sur ce sujet est

Rasmussen [2006]. Un GP est défini par la donnée de sa fonction moyenne et d'une fonction de covariance qui traduit la dépendance, en fonction de l'écart temporel ici, entre deux observations de F. Ainsi, k va définir le degré de lisse des trajectoires issues du processus ou autrement dit : Notons $T \subset \mathbb{R}^+$ un intervalle, $\forall \omega \in \Omega$, $t \in T$, la régularité de $\{Y_t(w), t \in T]\}$ dépend de k. On peut noter que toute fonction de covariance est nécessairement définie positive ce qui sera utile pour la suite. Grâce au théorème d'existence de Kolmogorov, on peut caractériser un processus gaussien par l'ensemble des n-uplets de variables aléatoires gaussiennes dont la matrice de variance-covariance est donnée par la valeur de la fonction de covariance entre chaque couple. C'est à dire, dire que $F \sim \mathcal{GP}(0, k)$ est équivalent à dire que $\forall n, \forall (t_1...t_n), (f(t_1)...f(t_n)) \sim \mathcal{N}(0, K)$ où K est défini comme suit :

$$K = \begin{pmatrix} k(t_1,t_1) & k(t_2,t_1) & \dots & k(t_n,t_1) \\ k(t_2,t_1) & . & & . \\ . & & & . \\ . & & . & . \\ k(t_n,t_1) & \dots & \dots & k(t_n,t_n) \end{pmatrix}$$

Car la fonction de covariance est symétrique. Dans le cas où la fonction de covariance k est totalement inconnue, le problème est très complexe. Il existe des méthodes d'estimation. On peut noter l'approche des moindres carrrés (Yao et al. 2005) dans le cas d'une régression selon une variable unidimensionnelle. Sinon on peut modéliser k par une somme de termes de covariances usuels qui peuvent introduire de la connaissance à priori sur le problème. Le choix de la fonction de covariance ou noyau est large. Parmi les noyaux usuels on notera l'exponentiel, le gaussien (exponentiel quadratique), le triangulaire ou le carré (boîte). En prenant $t_1, t_2 \in \mathbb{R}+$ et $r = |t_1 - t_2|$ la durée entre t_1 et t_2, les auteurs ont décomposé la covariance en termes deux termes.

La première composante représente la dépendance temporelle normale, d'un jour sur l'autre. Elle est modélisée par une exponentielle :

$$k_L(r) = \sigma_L^2 exp(-\frac{r^2}{2\delta_L^2})$$

La deuxième composante représente la dépendance de la réponse au rythme circadien :

$$k_S(r) = \sigma_S^2 exp(-\frac{r^2}{2\delta_S^2} - sin^2(\frac{2\pi r}{P_L}))$$

des mesures prises au même moment de la journée d'un jour sur l'autre sont plus corrélées que des mesures prises à 6h d'écart.

Un bon choix pour la forme de la fonction de covariance est cruciale, elle va dépendre des échelles de temps, du type de variables étudiées et peut faire l'objet d'une procédure de sélection de modèle. La figure 5 montre l'exemple des données de concentration en CO_2 en fonction des mois. Ici, le noyau non composite qui a été choisi ne parvient pas à capter la périodicité circadienne du niveau de dioxyde de carbone.

Inférence L'objectif de la RGP est d'estimer la fonction la plus probable F parmi l'ensemble de trajectoires issues d'un GP qui modélise notre série temporelle. Au vu des temps d'observation $(t_1, ..., t_n)$ de X, on a $(F(t_1)...F(t_n)) \sim \mathcal{N}(0, K)$. Par ailleurs, on peut déterminer simplement la loi de $(X(t_1)...X(t_n))$ par les propriétés des distributions gaussiennes.

Propriété 3. *Soit $f \sim \mathcal{N}(0_n, K)$ et soit $X = f + \epsilon$ où $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$, un vecteur gaussien. Alors X est également un vecteur gaussien et $X \sim \mathcal{N}(0_n, K + \sigma^2 I_n)$.*

Démonstration. Un vecteur est gaussien ssi toute combinaison linéaire de ses composantes est une variable gaussienne. Soit $a \in \mathbb{R}^n$, alors $a^T y = a^T f + a^T \epsilon$. Comme f, ϵ sont des vecteurs gaussiens, $a^T f$ et $a^T \epsilon$ sont des variables aléatoires gaussiennes, qui plus est indépendantes, alors $a^T X$ est gaussienne par somme. Par ailleurs, X est d'espérance nulle par linéarité : $E(X) = E(f) + E(\epsilon) = 0_n$. De plus, la

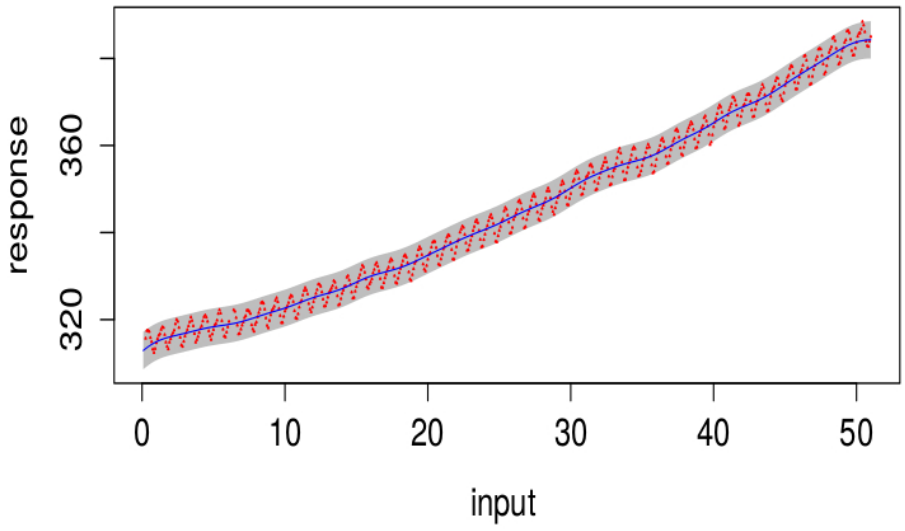

FIGURE 5 – Représentation de la densité prédictive issue d'une RGP sur les données de concentration atmosphérique mensuelles en dioxyde de carbone à Mauna Loa de 1960 à 2003. Illustration tirée de Rasmussen [2006].

matrice de covariance de X est $Cov(X,X) = Cov(f+\epsilon, f+\epsilon) = Cov(f,f) + Cov(\epsilon,\epsilon) + Cov(f,\epsilon) + Cov(\epsilon,f) = K + \sigma^2 I_n$ par bilinéarité de la covariance et indépendance de ϵ et f. □ □

On a donc :

$$X \sim \mathcal{N}(0_n, K + \sigma^2 I_n)$$

Ici, notons bien que la matrice de covariance $K + \sigma^2 I_n$ dépend du vecteur de paramètres θ. Notons $\mathcal{D} = (X,t)$ les données observées pour une série temporelle de la variable d'intérêt. Notons p la dimension de θ, c'est à dire le nombre de paramètres que l'on souhaite estimer. Pour un modèle et une forme de fonction de covariance donnée, le but de l'inférence dite inférence bayésienne empirique est de déterminer θ qui maximise vraisemblance conditionnelle des données $p(X|\theta)$ (les temps d'observations sont tacites sous cette notation). Les données étant fixées, cela définit une mesure sur l'espace $\mathbb{R}^p$ de définition de θ. On l'appelle vraisemblance marginale de θ, dont on prend le logarithme pour simplifier les calculs qui suivront. On le note $\mathcal{L}(\mathcal{D}|\theta)$. Notons $\hat{\theta}$ son estimateur, on a alors :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{Argmax}\{\mathcal{L}(\mathcal{D}|\theta)\}$$

Notons, $\Psi(\theta) := (K + \sigma^2 I_n)^{-1}$ qui dépend en effet de θ par la matrice K. D'après cette définition et la distribution gaussienne multivariée de $X|\theta$ on arrive très simplement à partir de la densité à l'expression suivante :

$$\mathcal{L}(\mathcal{D}|\theta) = -\frac{1}{2} log|\Psi(\theta)^{-1}| - -\frac{1}{2} y^T \Psi(\theta) y + cste$$

Pour présenter brièvement la méthode d'optimisation classique, elle consiste en un algorithme du gradient conjugué à partir du gradient et de la hessienne de la vraisemblance en tant que fonction de θ. La formule de celles-ci et donné pour tout paramètre individuel θ_i par :

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = -\frac{1}{2} tr\left(\Psi \frac{\partial \Psi^{-1}}{\partial \theta_i}\right) - \frac{1}{2} X^T \Psi \frac{\partial \Psi^{-1}}{\partial \theta_i} \Psi X$$

$$= tr\left((\alpha\alpha^T - \Psi^{-1})\frac{\partial\Psi}{\partial\theta_i}\right)$$

où tr signifie la trace de la matrice et $\alpha = \Psi X$. Ensuite, la matrice hessienne est $\forall j \in \{1...p\}$:

$$\frac{\partial\mathcal{L}}{\partial\theta_i\partial\theta_j} = \frac{1}{2}tr\left((\alpha\alpha^T - \Psi)(\frac{\partial^2\Psi^{-1}}{\partial\theta_i\partial\theta_j} - A_{ij}) - \alpha\alpha^T A_{ij}\right)$$

où

$$A_{ij} = \frac{\partial\Psi^{-1}}{\partial\theta_i}\Psi\frac{\partial\Psi^{-1}}{\partial\theta_j}$$

Un algorithme basé sur le gradient est un choix intéressant et peu coûteux computationnellement tant que l'inversion Ψ^{-1} est peu complexe. Cette complexité est en $O(n^3)$, où n est le nombre de points de la série temporelle. Dans notre cas, si l'on ajuste des GP indépendamment pour chaque patient et paramètre vital, la vitesse de calcul n'est pas une problématique majeure, car n est relativement faible. La question pourrait se poser si l'on essayait d'ajuster conjointement plusieurs signes vitaux en même temps qui ne soient pas indépendants, pour estimer par exemple leur interaction. Dans ce cas, la complexité augmenterait en $O((Dn)^3)$.

Il faut aussi noter qu'il que log-vraisemblance négative n'est pas convexe en général par rapport aux paramètres GP [Rasmussen, 2006]. Il existe potentiellement plusieurs minima locaux et globaux. Ces derniers peuvent incarner différents points de vue sur la fonction sous-jacente au vue des données, qui peuvent ne pas être évidents à trancher sans connaissance à priori du problème, comme le montre l'exemple de la figure 6.

Définition de la densité prédictive et comparaison de trajectoires Partant de séries temporelles très hétérogènes en instants d'observations, le processus gaussien choisi comme prior de la fonction du modèle va nous permettre d'estimer une fonction temporelle sous-jacente à la variable étudiée et surtout de quantifier l'incertitude autour de celle-ci.

D'après ce qui précède, on sait que $X \sim \mathcal{N}(0_n, K + \sigma^2 I_n)$ et l'on suppose que l'on a estimé par la méthode précédente des paramètres $\hat{\theta}$. On définit alors la densité prédictive associée à la régression selon ce modèle. $\forall t* \in \mathbb{R}+$, la densité prédictive au temps $t*$ est la loi de la variable aléatoire $X*$ d'une observation du signe vital à cet instant, conditionnellement aux observations à $(t_1, ..., t_n)$. Le processus et la variance sont fixés par $\hat{\theta}$. Cette loi s'exprime facilement grâce aux propriétés des distributions conditionnelles gaussiennes. En effet, si on s'intéresse au vecteur gaussien :

$$\begin{pmatrix} X(t_1) \\ . \\ . \\ . \\ X(t_n) \\ X* \end{pmatrix} \sim \mathcal{N}\left(0_{n+1}, \begin{pmatrix} K + \sigma^2 I_n & K* \\ K*^T & K** + \sigma^2 \end{pmatrix}\right)$$

où la matrice de covariadnce, découpées en blocs ici ($K \in \mathcal{M}_{n,n}, K* \in \mathcal{M}_{n,1}, K** \in \mathbb{R}$), est la matrice de variance covariance du vecteur gaussien $(X(t_1), ..., X(t_n), X*)$ définie par la fonction de covariance de paramètres $\hat{\theta}$. Alors, grâce à un théorème de distribution conditionnelle gaussienne la densité prédictive de $X*$ en $t*$ est :

$$X*|X \sim \mathcal{N}\left(K*^T (K + \sigma^2 I_n)^{-1} X \, , \, K** + \sigma^2 - K*^T (K + \sigma^2 I_n)^{-1} K*\right)$$

On peut alors utiliser la moyenne de la densité prédictive en tout instant comme estimation de la courbe vitale, corrigée des erreurs de mesure, au vu de nos données, et la variance comme indicateur de l'incertitude d'estimation en chaque instant du domaine d'observation.

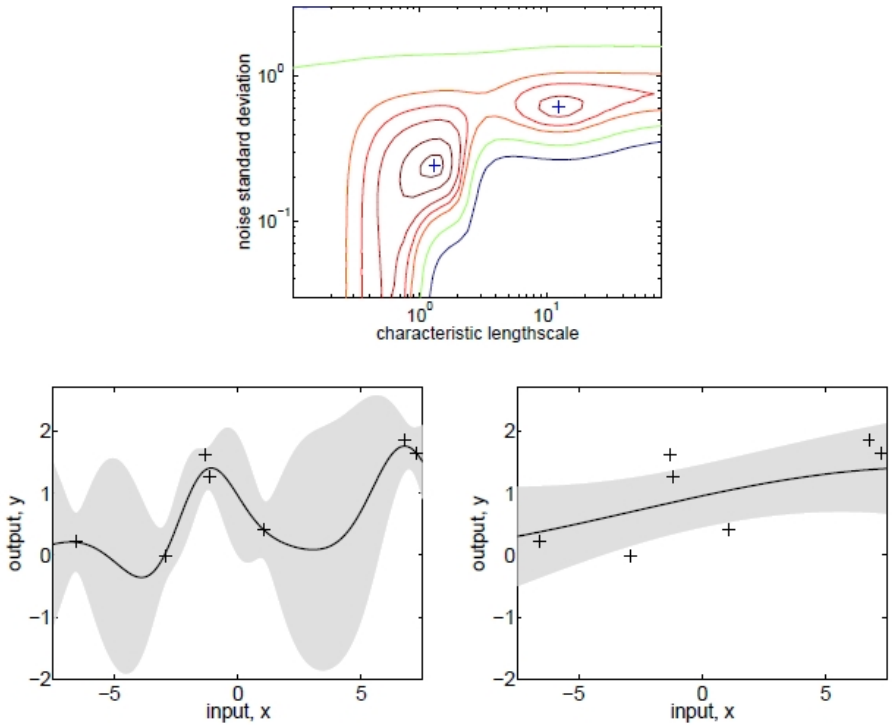

FIGURE 6 – Représentation d'une série temporelle (gauche et droite) et de la vraisemblance de celle-ci en fonction de deux paramètres (haut). La surface ombrée représente l'intervalle à 95% de la densité prédictive selon chaque estimation des paramètres aux optimas de la vraisemblance. Illustration tirée de Rasmussen [2006].

10.2 Intégration d'un processus gaussien

Rappel des hypothèses 1 et 2 :

1. $F \sim \mathcal{GP}(0_{\mathbb{R}^\tau}, k)$ est un processus gaussien de moyenne constante et nulle, et de fonction de covariance k.

2. $\zeta \sim \mathcal{N}(0_m, \gamma^2 I_m)$.

3. Il existe $g : \mathbb{R}^+ \to \mathbb{R}$ une application strictement monotone connue telle que $g(\tau)$ suit une loi normale. On note ainsi la transformation de temps de survie $Y := g(\tau)$, cette variable incarnera donc le label de notre modèle de survie.

4. $\beta \in \mathbb{R}^{\mathbb{R}^2}$ une fonction continue et bornée sur la surface $\{(s,t) \,/\, t \in \mathcal{T}, inf(\mathcal{T}) \leq s \leq t\}$.

5. $\epsilon \sim \mathcal{N}(0, \sigma^2), \sigma^2 \in \mathbb{R}^+$.

La **propriété 1** dit que :
Avec les définitions précédentes et les hypothèses ci-dessus on a :

$$\begin{pmatrix} \frac{1}{T}\int_0^T F(s)\beta(s,T)ds \\ X(t_1) \\ \vdots \\ X(t_m) \end{pmatrix} \sim \mathcal{N}\left(0_{m+1}, \begin{pmatrix} \frac{1}{T^2}\int_{[0,T]^2} k(s,t)\beta(s,T)\beta(t,T)dsdt & \frac{1}{T}\int_{[0,T]} K_*^T(s)\beta(s,T)ds \\ \frac{1}{T}\int_{[0,T]} \beta(s,T)K_*(s)ds & K_t + \gamma^2 I_m \end{pmatrix}\right)$$

Démonstration. Soit $T \in \mathbb{R}^{+*}$, $F \sim \mathcal{GP}(f,k)$ où $f \in \mathbb{R}^{\mathbb{R}}$ et $k \in \mathbb{R}^{\mathbb{R}^2}$ sont des fonctions continues bornées respectivement sur $[0,T]$ et $[0,T]^2$. Soit $\beta \in \mathbb{R}^{\mathbb{R}^2}$ une fonction continue bornée sur $[0,T]^2$. Soit $X = (X(t_1), ..., X(t_m))^T$ définit par le modèle **(1)**. On définit alors la suite de variables aléatoires réelles $(Y_n)_{n\in\mathbb{N}^*}$ par $Y_n = \sum_{j=1}^{n} \frac{T}{n}\beta\left(jT/n, T\right) F\left(jT/n\right) \; \forall n \in \mathbb{N}^*$.
D'après la caractérisation d'un processus gaussien, n-uplet de valeurs ponctuelles du processus gaussien F est un vecteur gaussien, en particulier $\forall n \in \mathbb{N}^*$, $F' = (F(T/n), ..., F(T), F(t_1), ..., F(t_m))^T$ est un vecteur gaussien. On peut donc construire par somme par somme des vecteur F' et $(0, ...0, \zeta)^T$, qui sont indépendants, le vecteur gaussien de taille $n + m$:

$$\begin{pmatrix} F(T/n) \\ \vdots \\ F(T) \\ X(t_1) \\ \vdots \\ X(t_m) \end{pmatrix} \sim \mathcal{N}\left(0_{m+n}, \begin{pmatrix} k(T/n,T/n) & & & & \\ \vdots & \ddots & & & \\ k(t_1,T/n) & & k(t_1,t_1)+\gamma & & \\ \vdots & & & \ddots & \\ k(t_m,T/n) & \cdots & & & k(t_m,t_m)+\gamma \end{pmatrix}\right)$$

A présent, on considère le vecteur $U_n := (Y_n, X(t_1), ..., X(t_m))^T$ qui est également gaussien car il est construit à partir de combinaisons linéaires des éléments du vecteur gaussien ci-dessus, et on peut expliciter son espérance et sa matrice de variance-covariance. Tout d'abord, Y_n est d'espérance et variance :

$$
\begin{aligned}
\mathbb{E}(Y_n) &= \frac{T}{n}\sum_{j=1}^{n}\beta(jT/n,T)\mathbb{E}(F(jT/n)) = \frac{T}{n}\sum_{j=1}^{n}\beta(jT/n,T)f(jT/n) \\
\mathbb{V}(Y_n) &= \mathbb{E}\left(\left(\frac{T}{n}\sum_{j=1}^{n}\beta(jT/n,T)\left(F(jT/n)-\mathbb{E}(F(jT/n))\right)\right)^2\right) \\
&= \frac{T^2}{n^2}\sum_{i,j\in[|1,n|]}\beta(iT/n,T)\,\beta(jT/n,T)\,\mathbb{C}ov\left(F(iT/n),F(jT/n)\right) \\
&= \frac{T^2}{n^2}\sum_{i,j\in[|1,n|]}\beta(iT/n,T)\,\beta(jT/n,T)\,k(iT/n,jT/n)
\end{aligned}
$$

Ensuite, $\forall k \in \{1,...,m\}$,

$$
\begin{aligned}
\mathbb{C}ov(Y_n, X(t_k)) &= \mathbb{C}ov\left(\tfrac{T}{n}\textstyle\sum_{j=1}^{n}\beta(jT/n,T)F(jT/n), F(t_k)+\zeta_k\right) \\
&= \tfrac{T}{n}\textstyle\sum_{j=1}^{n}\beta(jT/n,T)\mathbb{C}ov\left(F(jT/n),F(t_k)\right) \\
&= \tfrac{T}{n}\textstyle\sum_{j=1}^{n}\beta(jT/n,T)k(jT/n,t_k)
\end{aligned}
$$

car la covariance est une forme bilinéaire et par indépendance de l'erreur ζ_k et F. Ainsi, la matrice de variance-covariance de U_n est entièrement calculée par symétrie et de forme :

$$
\mathbb{V}(U_n) = \begin{pmatrix} \mathbb{V}(Y_n) & \mathbb{C}ov(X,Y_n)^T \\ \mathbb{C}ov(X,Y_n) & K_t+\gamma I_m \end{pmatrix}
$$

On cherche à présent montrer que la loi de la limite du vecteur U_n lorsque n tend vers l'infini est gaussienne. La fonction caractéristique de U_n est définie sur $\mathbb{R}^{m+1}$ par :

$$
\varphi_n : s \to \exp\left(s^T\mathbb{E}(U_n)i - \frac{1}{2}s^T\mathbb{V}(U_n)s\right)
$$

$\mathbb{V}(U_n)$ et $\mathbb{E}(U_n)$ admettent des limites finies lorsque $n \to +\infty$, car β et k et f sont des fonctions continues et bornées sur $[0,T]^2$. On note alors les limites suivantes :

$$
\begin{aligned}
\mathbb{E}(U) &= \lim_{n\to\infty}\mathbb{E}(U_n) = (\textstyle\int_0^T \beta(s,T)f(s)ds, 0, ...)^T \\
\mathbb{V}(U) &= \lim_{n\to\infty}\mathbb{V}(U_n) = \begin{pmatrix} \int_{[0,T]^2}\beta(s,T)\beta(x,T)k(s,x)dsdx & \int_{[0,T]}\beta(s,T)K_*^T(t,s)ds \\ \int_{[0,T]}\beta(s,T)K_*(t,s)ds & K_t+\gamma I_m \end{pmatrix}
\end{aligned}
$$

Soit $u \in \mathbb{R}$, alors par addition et composition des limites,

$$
\varphi_n(t) \underset{n\to+\infty}{\to} \exp\left(s^T\mathbb{E}(U)i - \frac{1}{2}s^T\mathbb{V}(U)s\right)
$$

Par conséquent, φ_n converge simplement vers φ. Or, la matrice $\mathbb{V}(U)$ est une matrice de variance-covariance d'un vecteur gaussien car elle est semi-définie positive. Donc, φ est la fonction caractéristique d'un vecteur gaussien U d'espérance $\mathbb{E}(U)$ et de matrice de variance-covariance $\mathbb{V}(U)$. Alors, d'après le théorème de convergence de Lévy, on conclut que $U_n \xrightarrow{\mathcal{L}} U$ et en notant Y la limite de Y_n, on note

$$
Y = \int_0^T F(s)\beta(s,T)ds \sim \mathcal{N}\left(\int_0^T \beta(s,T)f(s)ds\,,\, \int_{[0,T]^2}\beta(s,T)\beta(x,T)k(s,x)dsdx\right)
$$

La propriété est une conséquence directe lorsque $f = 0_{\mathbb{R}^{\mathbb{R}}}$. □

Références

Amarasingham, Ruben, Moore, Billy J, Tabak, Ying P, Drazner, Mark H, Clark, Christopher A, Zhang, Song, Reed, W Gary, Swanson, Timothy S, Ma, Ying, & Halm, Ethan A. 2010. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical care*, **48**(11), 981–988.

Baillie, Charles A, VanZandbergen, Christine, Tait, Gordon, Hanish, Asaf, Leas, Brian, French, Benjamin, William Hanson, C, Behta, Maryam, & Umscheid, Craig A. 2013. The readmission risk flag : Using the electronic health record to automatically identify patients at risk for 30-day readmission. *Journal of hospital medicine*, **8**(12), 689–695.

Barrett, James E, & Coolen, Anthony CC. 2013. Gaussian process regression for survival data with competing risks. *arXiv preprint arXiv :1312.1591*.

Brousseau, David C, Owens, Pamela L, Mosso, Andrew L, Panepinto, Julie A, & Steiner, Claudia A. 2010. Acute care utilization and rehospitalizations for sickle cell disease. *Jama*, **303**(13), 1288–1294.

Choi, Taeryon, & Schervish, Mark J. 2007. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, **98**(10), 1969–1987.

Cox, David R. 1992. Regression models and life-tables. *Pages 527–541 of : Breakthroughs in statistics*. Springer.

Doob, J Li. 1944. The elementary Gaussian processes. *The Annals of Mathematical Statistics*, **15**(3), 229–282.

Ferraty, Frédéric, & Vieu, Philippe. 2006. *Nonparametric functional data analysis : theory and practice*. Springer Science & Business Media.

Futoma, Joseph, Morris, Jonathan, & Lucas, Joseph. 2015. A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics*, **56**, 229–238.

Gellar, Jonathan E, Colantuoni, Elizabeth, Needham, Dale M, & Crainiceanu, Ciprian M. 2014. Variable-domain functional regression for modeling ICU data. *Journal of the American Statistical Association*, **109**(508), 1425–1439.

Goldsmith, Jeff, Bobb, Jennifer, Crainiceanu, Ciprian M, Caffo, Brian, & Reich, Daniel. 2012. Penalized functional regression. *Journal of Computational and Graphical Statistics*.

Klein, John P, & Moeschberger, Melvin L. 2005. *Survival analysis : techniques for censored and truncated data*. Springer Science & Business Media.

Medicare-Payment-Advisory. 2007. *Report to the Congress : promoting greater efficiency in Medicare*. (MedPAC).

Pimentel, M, Clifton, D, Clifton, Lei, & Tarassenko, Lionel. 2013. Modelling patient time-series data from electronic health records using Gaussian processes. *Pages 1–4 of : Advances in Neural Information Processing Systems : Workshop on Machine Learning for Clinical Data Analysis*.

Ramsay, James O. 2006. *Functional data analysis*. Wiley Online Library.

Rasmussen, Carl Edward. 2006. Gaussian processes for machine learning.

Wood, Simon N. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **65**(1), 95–114.

Wu, Lang, Liu, Wei, Yi, Grace Y, & Huang, Yangxin. 2011. Analysis of longitudinal and survival data : joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, **2012**.