Modélisation statistique et espaces de Hilbert autoreproduisants
Mini-cours

Anestis Antoniadis (UJF)
e-mail:antonia@imag.fr

Réunion de rentrée MAFIA - Nissan les Enserune, 2007

Plan du cours

- Motivation et introduction.
- Quelques rappels sur les espaces de Hilbert.
- Espaces hilbertiens à noyau auto-reproduisant (RKHS).
- Constructions
- Noyaux (semi-)définis positifs et RKHS.
- Exemples de RKHS

Plan du cours (Modélisation statistique)

- Régularisation en estimation non paramétrique.
- Le théorème du représentant.
- Les splines de lissage.
- Produits tensoriels et analyse de la variance fonctionnelle.
- Ondelettes et espaces auto-reproduisants.

Première Partie

Les espaces auto-reproduisants

Introduction

Lisser $\Leftrightarrow$ ôter la variabilité des données due à des causes non assignables pour découvrir des propriétés caractéristiques.

Depuis quelques années le terme lissage a pris une connotation spéciale en modélisation statistique et est devenu synonyme d' estimation non paramétrique et c'est dans ce sens que nous utiliserons ce terme dans ce cours.

La présentation dans ce cours sera axée sur le modèle de régression, mais la méthodologie est applicable à d'autres modèles (séries chronologiques, classification, etc . . .).

Modèle

Le scénario sera le suivant : des réalisations de $Y_1, \ldots, Y_n$ sont observées en des points $x_1, \ldots, x_n$, selon le modèle

$$Y_j = r(x_j) + \epsilon_j, \quad j = 1, \ldots, n, \quad \text{où}$$

- r est une fonction définie sur $\mathcal{T}$
- les $x_i \in \mathcal{T}$ et
- $\epsilon_1, \ldots, \epsilon_n$ sont des v.a. non observées représentant les erreurs.

L'objectif est d'identifier le plus raisonnablement r et une des méthodes les plus courantes est à l'aide des fonctions splines de lissage et de leur représentation dans les RKHS.

Les espaces de Hilbert à noyau autoreproduisant

– Analyse (sys. orth. de fonctions harmoniques (Bergman (1921)) – Esp. fonctionnels à noyau reproduisant (Aronszajn (1943-1951)) – Sous-espace hilbertiens d'e.v.t (Schwartz (1962)) – . . .).

– Statistique (Stat. des processus de carré intégrable (Parzen (1958–1970)) – Plans d'expérience en régression (Ylvisaker et Sacks (1961)) – Splines de lissage (Wahba (1971–1990)).

– Probabilité (Mesures gaussiennes sur espaces fonctionnels (Kuelbs (1968)) - proc. centrés et RKHS (Loeve (1943)) .

– Traitement du signal (Filtrage de Kalman (Laning et Batton (1957)) - FDA (Ramsay, Silverman) – méthodes à noyau (SVM, . . .).

Rappels sur les espaces de Hilbert

Les espaces de Hilbert que nous allons considérer dans la suite seront des espaces vectoriels sur le corps des réels $\mathbb{R}$ ou le corps des complexes $\mathbb{C}$.

Rappelons qu'un produit scalaire sur un espace vectoriel $\mathcal{H}$ sur $\mathbb{R}$ est une application $(f, g) \rightarrow \langle f, g \rangle_{\mathcal{H}}$ de $\mathcal{H}^{\times 2}$ dans $\mathbb{R}$ qui est bilinéaire, symétrique et définie positive (i.e. $\langle f, f \rangle_{\mathcal{H}} > 0$ pour tout $f \in \mathcal{H} \backslash \{0\}$).

Un espace vectoriel muni d'un produit scalaire est appelé pré-hilbertien. Il est muni d'une norme associée au produit scalaire par $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$.

Un **espace de Hilbert** est un espace vectoriel muni d'un produit scalaire et complet pour la norme associée.

Rappels (suite)

Un sous-espace fermé d'un espace de Hilbert est lui-même un espace de Hilbert. La distance d'un élément $f \in \mathcal{H}$ à un sous-espace vectoriel fermé $\mathcal{G} \subset \mathcal{H}$ est définie par

$$D(f, \mathcal{G}) = \inf_{g \in \mathcal{G}} \|f - g\|_{\mathcal{H}}.$$

Du fait que $\mathcal{G}$ est fermé, il existe un unique $f_{\mathcal{G}} \in \mathcal{G}$, la projection de f sur $\mathcal{G}$ telle que $D(f, \mathcal{G}) = \|f - f_{\mathcal{G}}\|_{\mathcal{H}}$. De plus,

$$\langle f - f_{\mathcal{G}}, g \rangle_{\mathcal{H}} = 0, \quad \forall g \in \mathcal{G}.$$

Cela entraîne que

$$\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^{\perp},$$

où $\mathcal{G}^{\perp} = \{g \in \mathcal{H}; \langle f, g \rangle_{\mathcal{H}} = 0, \quad \forall g \in \mathcal{G}\}$.

Rappels (suite)

Si $\mathcal{H}_1$ et $\mathcal{H}_2$ sont deux espaces de Hilbert, sous-espaces d'un espace vectoriel $\mathcal{L}$ n'ayant comme élément commun que $0 \in \mathcal{L}$, alors l'espace $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ avec pour éléments $f = f_1 + f_2$ ($f_1 \in \mathcal{H}_1$, $f_2 \in \mathcal{H}_2$) et $g = g_1 + g_2$ ($g_1 \in \mathcal{H}_1$, $g_2 \in \mathcal{H}_2$) et produit scalaire $\langle f, g \rangle_{\mathcal{H}} = \langle f_1, g_1 \rangle_{\mathcal{H}_1} + \langle f_2, g_2 \rangle_{\mathcal{H}_2}$ est un espace de Hilbert.

Rappelons enfin le **théorème de représentation de Riesz** :
Pour toute forme linéaire continue L sur un espace de Hilbert $\mathcal{H}$ il existe un unique élement g_L dans $\mathcal{H}$ tel que $\forall f \in \mathcal{H}$, $\langle g_L, f \rangle_{\mathcal{H}} = Lf$.

Notons N_L le noyau de L. Comme L est continue, N_L est fermé. Si $N_L = \mathcal{H}$ on prend $g_L = 0$; sinon, il existe un élément non nul $g_0 \in N_L^{\perp}$. Il est facile de voir que $N_L^{\perp}$ est unidimensionnel ($f - (Lf)g_0/Lg_0 \in N_L$ et si $f \in N_L^{\perp}$ alors $f = (Lf)g_0/Lg_0$). On prend donc $g_L = (Lg_0)g_0/\|g_0\|^2_{\mathcal{H}}$, l'unicité étant évidente.

RKHS: le cas de la dimension finie

Soit $\mathcal{T} = \{1, 2, \ldots, n\}$ etnotons $E = \mathbb{R}^{\mathcal{T}}$ l'espace des fonctions numériques définies sur $\mathcal{T}$ qui, vue la finitude de $\mathcal{T}$, est isomorphe à $\mathbb{R}^n$.

soit $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,n}$ une matrice carrée d'ordre n définie strictement positive. La matrice Σ permet de définir sur E une structure d'espace Hilbertien en prenant pour produit scalaire

$$\langle f, g \rangle_E = \mathbf{f}^T \Sigma^{-1} \mathbf{g},$$

où $\mathbf{f}^T = (f(1), \ldots, f(n))$. Notons $\boldsymbol{\sigma}_i$, $i = 1, \ldots, n$ les colonnes de Σ. On constate que l'on a

$$\langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle_E = \sigma_{ij}.$$

Cas de la dimension finie

Pourquoi $\langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle_E = \sigma_{ij}$? On a

$$\langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle_E = \boldsymbol{\sigma}_i^T \Sigma^{-1} \boldsymbol{\sigma}_j = \boldsymbol{\sigma}_i^T \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \sigma_{ij},$$

car $\Sigma^{-1}\Sigma = \mathbf{I}_n$. Plus généralement, pour $\mathbf{f} = (f(1), \ldots, f(n))^T$, on a $\langle \boldsymbol{\sigma}_i, \mathbf{f} \rangle_E = f(i)$.

On constate donc que les fonctionnelles d'évaluation $i \to f(i)$ sont des éléments de E, représentés par $\boldsymbol{\sigma}_i$, qui comme formes linéaires sur E de dimension finie sont continues. De plus, comme Σ est définie positive, les vecteurs $\boldsymbol{\sigma}_i$, $i = 1, \ldots, n$ engendrent E.

RKHS: le cas général

Soit $\mathcal{T}$ un ensemble quelconque et notons E l'espace vectoriel réels des fonctions définies sur $\mathcal{T}$ et à valeurs dans $\mathbb{R}$ (tout élément f de E est une fonction f définie sur $\mathcal{T}$ à valeurs réelles). On a :

Définition *On dit qu'un sous-espace $\mathcal{H} \subset E$ est un* **espace de Hilbert auto-reproduisant** *sur $\mathcal{T}$, si et seulement si :*

- *$\mathcal{H}$ est un sous-espace vectoriel de E*
- *Il est possible d'associer à $\mathcal{H}$ un produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ qui soit tel que $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ soit un espace de Hilbert.*
- *Pour tout $t \in \mathcal{T}$, la fonctionnelle linéaire d'évaluation $\delta_t : \mathcal{H} \to \mathbb{R}$, définie par $d_t(f) = f(t)$ est bornée (i.e l'injection de $\mathcal{H}$ dans E est continue).*

Le cas général (suite)

D'après la définition l'espace de Hilbert $\mathcal{H}$ est un RKHS si et seulement si pour tout $f \in \mathcal{H}$ et pour tout $t \in \mathcal{T}$, il existe M_t ne dépendant pas de f tel que $|f(t)| \leq M_t \|f\|_{\mathcal{H}}$. D'après le théorème de représentation de Riesz, il existe donc un unique représentant ξ_t dans $\mathcal{H}$ associé à δ_t, i.e. tel que, pour tout $f \in \mathcal{H}$ on ait $\langle \xi_t, f \rangle_{\mathcal{H}} = f(t)$.

Définition *La fonction* $t \in \mathcal{T} \rightarrow \xi_t \in \mathcal{H}$ *s'appelle fonction de reproduction en* t*, et la fonction de* $\mathcal{T}^{\times 2}$ *à valeurs dans* $\mathbb{R}$ *définie par*

$$K(s,t) = \xi_s(t),$$

est appelée le **noyau auto-reproduisant de** $\mathcal{H}$*.*
On remarquera que $\langle K(s,\cdot), K(t,\cdot) \rangle_{\mathcal{H}} = K(s,t)$ et $\|\xi_t\|^2_{\mathcal{H}} = K(t,t)$.

Propriétés

- Si $\mathcal{H}$ est un RKHS sur $\mathcal{T}$ son noyau est unique.
- Un noyau auto-reproduisant est semi-défini positif. Si $\mathcal{H}$ est séparant alors le noyau est défini positif.
- Toute suite de fonctions $\{f_n\}$ de $\mathcal{H}$ qui converge vers $f \in \mathcal{H}$ au sens de la norme de $\mathcal{H}$ converge aussi ponctuellement en tout point de $\mathcal{T}$.
- Supposons $\mathcal{H}$ un sous espace fermé d'un espace de Hilbert $(\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$ de fonctions définies sur $\mathcal{T}$ et à valeurs réelles et que le produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ est la trace sur $\mathcal{H}$ du produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{X}}$. Alors pour toute fonction $g \in \mathcal{X}$, la fonction $t \to \langle g, K(\cdot, t) \rangle_{\mathcal{X}}$ est la projection de g sur $\mathcal{H}$.

Preuves

• Supposons qu'il existe un autre noyau auto-reproduisant K'. Alors, pour tout $s \in \mathcal{T}$, nous avons

$$0 \leq \|(K(\cdot, s) - K'(\cdot, s)\|_{\mathcal{H}} = K(s, s) - K'(s, s) - K(s, s) + K'(s, s) = 0.$$

• Un noyau auto-reproduisant est symétrique car, pour tout $s, t \in \mathcal{T}$

$$K(s, t) = \langle K(s, \cdot), K(t, \cdot)\rangle_{\mathcal{H}} = \langle K(t, \cdot), K(s, \cdot)\rangle_{\mathcal{H}} = K(t, s).$$

Il est semi-défini positif car pour tout N, tout $\beta_1, \ldots, \beta_N \in \mathbb{R}$ et tout $t_1, \ldots, t_N \in \mathcal{T}$

$$\sum_{i,j=1}^{N} \beta_i \beta_j K(t_i, t_j) = \sum_{i,j=1}^{N} \beta_i \beta_j \langle K(t_i, \cdot), K(t_j, \cdot)\rangle_{\mathcal{H}} = \| \sum_{i=1}^{N} \beta_i K(t_i, \cdot)\|^2_{\mathcal{H}}.$$

Preuves (suite)

• La propriété d'auto-reproduction implique que, pour tout $t \in \mathcal{T}$
$f_n(t) - f(t)| = |\langle f_n - f, K(t,\cdot)\rangle_{\mathcal{H}} \leq \|f_n - f\|_{\mathcal{H}} \|K(t,\cdot)\|_{\mathcal{H}} = K(t,t)^{1/2}\|f_n - f\|_{\mathcal{H}},$
et donc la convergence forte en norme $\|\cdot\|_{\mathcal{H}}$ implique la convergence ponctuelle en tout point.

• Pour la projection, on doit montrer que

1. La fonction $t \to \langle g, K(\cdot,t)\rangle_{\mathcal{X}}$ est dans $\mathcal{H}$
2. $(g - \langle g, K(\cdot,t)\rangle_{\mathcal{X}}) \perp \mathcal{H}$

Le théorème des projections montre que pour tout $g \in \mathcal{X}$, il existe un unique élement f de $\mathcal{H}$ tel que $g - f$ soit orthogonal à $\mathcal{H}$. Comme pour tout $t \in \mathcal{T}$, $t \to K(\cdot,t)$ est une fonction de $\mathcal{H}$, on a :

$$\langle g - f, K(\cdot,t)\rangle_{\mathcal{X}} = 0$$

et le résultat découle de la propriété d'auto-reproduction.

Le Théorème de Moore-Aronszajn

Nous venons de voir que le noyau auto-reproduisant K d'un RKHS $\mathcal{H}$ est un noyau défini positif sur $\mathcal{T}^{\times 2}$. Le théorème suivant établit la réciproque et permet de construire un RKHS à partir d'un noyau défini positif.

Théorème d'Aronszajn. *Soit $\mathcal{T}$ un ensemble quelconque. A tout noyau K défini positif sur $\mathcal{T} \times \mathcal{T}$ correspond un unique espace hilbertien reproduisant $\mathcal{H}_K$ de fonctions sur $\mathcal{T}$ dont le noyau autoreproduisant est K. Pour tout $f \in \mathcal{H}_K$ et tout $t \in \mathcal{T}$ on a $\langle K(t, \cdot), f \rangle_{\mathcal{H}_K} = f(t)$.*

Construction dans le cas où $\mathcal{T}$ est fini

On suppose que $\mathcal{T} = \{t_1, \ldots, t_N\}$ est fini. La donnée d'un noyau défini positif sur $\mathcal{T}$ est alors équivalente à la donnée d'une matrice carrée symétrique d'ordre N, Σ, définie positive. Cette dernière est donc diagonalisable dans une base orthonormée $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ avec des valeurs propres $0 < \lambda_1 \leq \cdots \leq \lambda_N$ ce qui s'écrit

$$\sigma_{ij} = \sum_{k=1}^{N} \lambda_k u_{ik} u_{jk} = \sum_{k=1}^{N} \lambda_k \Phi_k(i) \Phi_k(j), \text{ avec } \Phi_k(i) = u_{ik}.$$

On voit aisément que

$$\langle \mathbf{f}, \mathbf{g} \rangle_E = \mathbf{f}^T \Sigma^{-1} \mathbf{g} = \sum_{k=1}^{N} \frac{\langle \mathbf{f}, \mathbf{\Phi}_k \rangle_N \langle \mathbf{g}, \mathbf{\Phi}_k \rangle_N}{\lambda_k},$$

où $\langle \cdot, \cdot \rangle_N$ est le produit euclidien classique de $\mathbb{R}^N$.

Construction dans le cas où $\mathcal{T}$ est compact et K est continu

Considérons maintenant le cas de $\mathcal{T}$ un espace métrique compact (typiquement un fermé borné dans $\mathbb{R}^d$) et soit K un noyau défini positif *continu* sur $\mathcal{T} \times \mathcal{T}$.

Un tel noyau est appelé *Noyau de Mercer*.

La construction de l'espace auto-reproduisant associé tente de mimer ce qui se passe dans le cas fini et repose sur une série de lemmes que nous allons examiner.

Noyau de Mercer

Soit μ une mesure Borélienne sur $\mathcal{T}$ et $\mathcal{H} = L_2(\mathcal{T}, d\mu)$.

Pour toute fonction $K : \mathcal{T}^{\times 2} \to \mathbb{R}$ on pose (lorsque c'est défini):

$$(L_K f)(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d\mu(\mathbf{t}).$$

On a alors:

Lemme 1 Si K est un noyau de Mercer, alors L_K est un opérateur linéaire borné compact de $L_2(\mathcal{T}, d\mu)$, auto-adjoint et positif.

Rappels

Soit $\mathcal{H}$ un espace hilbertien.

- Un *opérateur linéaire* L sur $\mathcal{H}$ est une application linéaire continue de $\mathcal{H}$ dans lui-même.
- On dit que L est *compact* si pour toute suite bornée $\{u_n\}_{n=1}^{\infty}$ de $\mathcal{H}$, la suite Lu_n possède une sous-suite convergente.
- L est *auto-adjoint* si pour tout $f, g \in \mathcal{H}$ on a $\langle f, Lg\rangle = \langle Lf, g\rangle$
- L est positif ssi il est auto-adjoint et pour tout $f \in \mathcal{H}$ on a $\langle f, Lf\rangle \geq 0$.

Preuve du lemme 1

- L_K est une application linéaire de $L_2(\mathcal{T}, d\mu)$ dans $L_2(\mathcal{T}, d\mu)$
- L_K est borné
- L_K est compact (Ascoli)
- L_K est auto-adjoint
- L_K est positif

L_K est linéaire de $L_2(\mathcal{T}, d\mu)$ dans $L_2(\mathcal{T}, d\mu)$

Pour tout $f \in L_2(\mathcal{T}, d\mu)$ et tout $(\mathbf{t}_1, \mathbf{t}_2) \in \mathcal{T}^{\times 2}$ on a :

$$\begin{aligned} |L_K f(\mathbf{t}_1) - L_K f(\mathbf{t}_2)| &= \left| \int (K(\mathbf{t}_1, \mathbf{x}) - K(\mathbf{t}_2, \mathbf{x}))\, f(\mathbf{x}) d\mu(\mathbf{x}) \right| \\ &\leq \|K(\mathbf{t}_1, \cdot) - K(\mathbf{t}_2, \cdot)\|_2 \|f\|_2 \quad \text{(Cauchy-Schwartz)} \\ &\leq \max_{\mathbf{x} \mathcal{T}} |K(\mathbf{t}_1, \mathbf{x}) - K(\mathbf{t}_2, \mathbf{x})| \|f\|_2 \sqrt{\mu(\mathcal{T})}. \end{aligned}$$

K étant continu et $\mathcal{T}$ compact, K est uniformément continu et donc $L_K f$ est continue, donc de carré intégrable.

L_K est borné

Pour tout $f \in L_2(\mathcal{T}, d\mu)$ et tout $\mathbf{x} \in \mathcal{T}$ on a :

$$|L_K f(\mathbf{x})| = \left| \int K(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d\mu(\mathbf{t}) \right| \leq \sqrt{\mu(\mathcal{T})} C_{K,x} \|f\|_2,$$

où $C_{K,x} = \max_{\mathbf{t} \in \mathcal{T}} |K(\mathbf{t}, \mathbf{x})|$. Donc

$$\|L_K f\|_2 = \left(\int L_K f(\mathbf{x})^2 d\mu(\mathbf{x}) \right)^{1/2} \leq C_K \mu(\mathcal{T}) \|f\|_2.$$

Rappel: Théorème d'Ascoli

Soit $C(\mathcal{T})$ l'ensemble des fonctions continues sur $\mathcal{T}$, muni de la norme sup. Un ensemble de fonctions $G \subset C(\mathcal{T})$ est dit *équicontinu* ssi :

$\forall \epsilon > 0, \exists \delta > 0, \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{T},$

$$\|\mathbf{x}_1 - \mathbf{x}_2\| < \delta \Rightarrow \forall g \in G, |g(\mathbf{x}_1) - g(\mathbf{x}_2)| < \epsilon.$$

Théorème 6 (Ascoli) Une partie $H \subset C(\mathcal{T})$ est relativement compacte (i.e. son adhérence est compacte) ssi elle est uniformément bornée et équicontinue.

L_K est compact

Soit $\{f_n\}_{n\geq 0}$ une suite bornée de $L_2(\mathcal{T}, d\mu)$. La suite $\{L_K f_n\}_{n\geq 0}$ est une suite de fonctions continues, uniformément bornée car

$$\|L_K f\|_\infty \leq \sqrt{\mu(\mathcal{T})} C_K \|f\|_2 \leq \mu(\mathcal{T}) C_K M.$$

D'autre part elle est équicontinue, car

$$|L_K f_n(\mathbf{x}_1) - L_K f_n(\mathbf{x}_2)| \leq \sqrt{\mu(\mathcal{T})} \max_{\mathbf{t}\in\mathcal{T}} |K(\mathbf{x}_1, \mathbf{t}) - K(\mathbf{x}_2, \mathbf{t})| M.$$

Selon le théorème d'Ascoli, on peut donc extraire une sous-suite uniformément convergente dans $C(\mathcal{T})$, et donc dans $L_2(\mathcal{T}, d\mu)$.

L_K est auto-adjoint et positif

K étant symétrique on a pour tout $f, g \in L_2(\mathcal{T}, d\mu)$:

$$\langle f, L_K g \rangle = \int f(\mathbf{x}) L_K g(\mathbf{x}) d\mu(\mathbf{x}) =$$

$$\int f(\mathbf{x}) g(\mathbf{t}) K(\mathbf{x}, \mathbf{t}) d\mu(\mathbf{x}) d\mu(\mathbf{t}) \text{ Fubini} = \langle L_K f, g \rangle$$

$$\langle f, L_K f \rangle = \int f(\mathbf{x}) f(\mathbf{t}) K(\mathbf{x}, \mathbf{t}) d\mu(\mathbf{x}) d\mu(\mathbf{t}) =$$

$$\lim_{k \to \infty} \frac{\mu(\mathcal{T})}{k^2} \sum_{i,j=1}^{k} K(\mathbf{x}_i, \mathbf{x}_j) f(\mathbf{x}_i) f(\mathbf{x}_j)$$

Théorème spectral

Lemme 2 *Soit L un opérateur linéaire compact sur un espace de Hilbert $\mathcal{H}$. Alors il existe dans $\mathcal{H}$ un système orthonormal complet $\{\psi_1, \psi_2, \dots\}$ de vecteurs propres de L. Les valeurs propres $\{\lambda_1, \lambda_2, \dots\}$ sont réelles si L est auto-adjoint, et positives si L est positif.*

Dans le cas de L_K, les vecteurs propres ψ_k associés aux valeurs propres $\lambda_k \neq 0$ sont des fonctions continues car :

$$\psi_k = \frac{1}{\lambda_k} L_K \psi_k.$$

Théorème de Mercer

Lemme 3 *Soit $\mathcal{T}$ un espace normé compact, μ une mesure borélienne sur $\mathcal{T}$ et K un noyau de Mercer. Soit $\{\lambda_1, \lambda_2, \dots\}$ les valeurs propres de L_K(rangées par ordre décroissant) et $\{\psi_1, \psi_2, \dots\}$ le système orthonormal complet de vecteurs propres associés. Alors, pour tout $\mathbf{x}, \mathbf{x}'$ dans $\mathcal{T}$ on a :*

$$K(\mathbf{x}, \mathbf{x}') = \sum_k \lambda_k \psi_k(\mathbf{x}) \psi_k(\mathbf{x}'),$$

la convergence étant absolue et uniforme sur $\mathcal{T}^{\times 2}$

De ce lemme on en déduit que $\Phi : \mathcal{T} \to \ell^2$ définie par $\Phi(\mathbf{x}) = \left\{\sqrt{\lambda_k}\psi_k(\mathbf{x})\right\}$ est bien définie, continue et telle que

$$K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\ell^2}.$$

Preuve

En effet, par le théorème de Mercer on voit que pour tout $\mathbf{x} \in \mathcal{T}$, $\sum_k \lambda_k \psi_k(\mathbf{x})^2$ converge vers $K(\mathbf{x}, \mathbf{x}) < \infty$, et donc $\Phi(\mathbf{x}) \in \ell_2$.

La continuité de Φ découle de :

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|_{\ell_2} = \sum_k \lambda_k (\psi_k(\mathbf{x}) - \psi_k(\mathbf{x}'))^2 =$$
$$= K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}')$$

Espace RKHS associé

Nous supposons que $\lambda_k > 0$ pour tout $k \geq 1$ (sinon, le résultat et la preuve restent valides dans le sous-espace engendré par les vecteurs propres de valeur propres non nulles). Soit l'espace de Hilbert :

$$H_K = \left\{ f \in L_2(\mathcal{T}, d\mu); f = \sum_{i=1}^{n} a_i \psi_i, \sum \frac{a_k^2}{\lambda_k} < \infty \right\}$$

muni du produit scalaire $\langle f, g \rangle_K = \sum_{k=1}^{\infty} \frac{a_k b_k}{\lambda_k}$.

Donc

$$\langle f, g \rangle = \sum_{k=1}^{\infty} \frac{\langle f, \psi_k \rangle_{L_2(\mathcal{T}, d\mu)} \langle g, \psi_k \rangle_{L_2(\mathcal{T}, d\mu)}}{\lambda_k},$$

et H_K est le rkhs associé au noyau K.

Le cas général : Théorème d'Aronszajn

Théorème *A chaque noyau semi-défini positif sur un ensemble arbitraire $\mathcal{T}$, $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, on peut associer un unique espace hilbertien réel $\mathcal{H}$ de fonctions, $\mathcal{H} \subset \mathbb{R}^{\mathcal{T}}$, cet espace admettant k comme noyau reproduisant.*

Notons que l'ensemble $\mathcal{T}$ est supposé être quelconque.

La preuve se fera en 2 étapes que nous allons détailler.

Première étape

On considère d'abord l'espace vectoriel $\mathcal{H}_0$ formé des combinaisons linéaires finies du type

$$f = \sum_{i=1}^{n} \alpha_i K(\cdot, \mathbf{x}_i), \quad \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{T}.$$

Pour $\mathbf{x}$ et $\mathbf{y} \in \mathcal{T}$, nous posons

$$K(\mathbf{x}, \mathbf{y}) = \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y}) \rangle_{\mathcal{H}_0},$$

produit que nous étendons à toutes fonctions $f = \sum_{i=1}^{n} \alpha_i K(\cdot, \mathbf{x}_i) \in \mathcal{H}_0$ et $g = \sum_{i=1}^{m} \beta_i K(\cdot, \mathbf{y}_i) \in \mathcal{H}_0$ par multi-linéarité :

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{y}_j).$$

Première étape (suite)

On a

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{j=1}^{m} \beta_j \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{y}_j) = \sum_{j=1}^{m} \beta_j \sum_{i=1}^{n} \alpha_i K(\mathbf{y}_j, \mathbf{x}_i) = \sum_{j=1}^{m} \beta_j f(\mathbf{y}_j)$$

expression qui ne dépend donc pas de la représentation de f dans $\mathcal{H}_0$, mais seulement de ses valeurs. Idem par symétrie pour g. Il est immédiat de prouver que la forme $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ est bilinéaire et positive. De plus si $\langle f, f\rangle_{\mathcal{H}_0} = 0$, on a, pour tout $\mathbf{x} \in \mathcal{T}$:

$$|f(\mathbf{x})| = \langle f, K(\cdot, \mathbf{x})\rangle_{\mathcal{H}_0} \leq K^{1/2}(\mathbf{x}, \mathbf{x}) \|f\|_{\mathcal{H}_0}$$

et donc $\|f\|_{\mathcal{H}_0} = 0$ implique que $f(\mathbf{x}) = 0$. Donc $\mathcal{H}_0$ est muni d'une structure d'espace préhilbertien.

Deuxième étape: Complétion de $\mathcal{H}_0$

Définition (Complétion fonctionnelle). *Soit $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ un espace pré-hilbertien de fonctions définies sur $\mathcal{T}$ et à valeurs réelles. On dit qu'il existe une complétion fonctionnelle de $\mathcal{H}_0$ si il existe*

- *un espace vectoriel de fonctions $\mathcal{H}$ définies sur $\mathcal{T}$ et à valeurs réelles, tel le que $\mathcal{H}_0 \subset \mathcal{H}$,*

- *un produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ sur $\mathcal{H}$*

tels que $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ soit un espace de Hilbert et , pour tout $\mathbf{x} \in \mathcal{T}$, la forme linéaire $\delta_{\mathbf{x}}(f) = f(\mathbf{x})$ est continue.

Complétion de $\mathcal{H}_0$ (suite)

Lemme. *Soit $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ un espace pré-hilbertien de fonctions définies sur $\mathcal{T}$ et à valeurs réelles. Il existe une complétion fonctionnelle de $\mathcal{H}_0$ si et seulement si*

- *pour tout $\mathbf{x} \in \mathcal{T}$, la forme linéaire $\delta_{\mathbf{x}}(f) = f(\mathbf{x})$ est bornée sur $\mathcal{H}_0$.*

- *pour toute suite de Cauchy $\{f_n\}_{n \geq 0}$ d'éléments de $\mathcal{H}_0$, la condition $\lim_{n \to \infty} f_n(\mathbf{x}) = 0$ pour tout $\mathbf{x} \in \mathcal{T}$ implique que $\|f_n\|_{\mathcal{H}_0} \to 0$.*

Si la complétion fonctionnelle existe, elle est unique.

La démonstration de l'étape 2 du théorème consiste alors à vérifier les hypothèses du lemme ci-dessus pour notre construction.

Complétion de $\mathcal{H}_0$ (suite)

On a, pour $f \in \mathcal{H}_0$ et $\mathbf{x} \in \mathcal{T}$,

$$|f(\mathbf{x})| = \langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_0} \leq K^{1/2}(\mathbf{x}, \mathbf{x}) \|f\|_{\mathcal{H}_0}$$

et donc $\delta_{\mathbf{x}}$ est bornée sur $\mathcal{H}_0$.

Soit maintenant une suite de Cauchy $\{f_n\}_{n \geq 0}$ d'éléments de $\mathcal{H}_0$ telle que $\lim_{n \to \infty} f_n(\mathbf{x}) = 0$ pour tout $\mathbf{x} \in \mathcal{T}$. On a

$$\|f_m\|^2 \leq \|f_n - f_m\|^2 + 2|\langle f_n, f_m \rangle|,$$

et par définition $\langle f_n, f_m \rangle = \sum_{i=1}^{p} \alpha_{n,i} f_m(\mathbf{x}_{n,i})$ ce qui implique que $\|f_m\|^2 \leq \lim_{n \to \infty} \|f_n - f_m\|^2$. Il est alors aisé de voir que K est le noyau reproduisant de $\mathcal{H}$.

Exemples de RKHS

Nous allons considérer ici quelques exemples de RKHS le plus fréquemment utilisés en statistique.

Commençons d'abord par un espace de Hilbert qui n'est pas un RKHS, l'espace $L^2([0,1])$. On considère donc l'espace des fonctions continues sur $[0,1]$, $C([0,1])$, sur lequel on définit la norme pré-hilbertienne $\|f\|^2 = \int_0^1 f(t)^2 dt$, que l'on complète pour obtenir l'espace de Hilbert $L^2([0,1])$.

Maintenant, pour tout point $x \in [0,1]$, il est facile de construire une suite $f_n \in C([0,1])$ telle que $\lim_n \|f_n\| = 0$ et $\lim_n f_n(x) = +\infty$.

Il est impossible de considérer donc les élements de $L^2([0,1])$ comme des fonctions, en particulier, cet espace de Hilbert n'est pas un RKHS!

RKHS et régularité fonctionnelle

Considérons l'espace

$$\mathcal{H} = \left\{ f \in [0,1] \to \mathbb{R}, \text{ abs. cont., dérivable p.p. , } f' \in L^2([0,1]),\ f(0) = 0 \right\}.$$

C'est un espace de Hilbert lorsqu'on le munit du produit scalaire

$$\langle f, g \rangle_{\mathcal{H}} = \int_0^1 f'(t) g'(t) dt,$$

et la norme $\|f\|_{\mathcal{H}}$ mesure la régularité de f. En fait, $\mathcal{H}$ est un RKHS sur $\mathcal{T} = [0,1]$ avec pour noyau K la fonction définie positive

$$K(s,t) = \min(s,t)$$

Preuve

On a, pour tout $x \in [0,1]$ et $f \in \mathcal{H}$,

$$|f(x)| = |f(x) - 0| = |f(x) - f(0)| = |\int_0^x f'(t)dt| \leq \sqrt{(x)}\|f\|_{\mathcal{H}},$$

et donc la fonctionnelle d'évaluation est bien définie et continue sur $\mathcal{H}$.

Par définition, pour tout $x \in [0,1]$, $K(\cdot, x) = \min(\cdot, x)$ est continue avec $K(0,x) = 0$, dérivable (sauf en x), avec un dérivée de carré intégrable et donc $K(\cdot, x) \in \mathcal{H}$. De plus

$$\langle f, K_x \rangle_{\mathcal{H}} = \int_0^1 f'(t)K'_x(t)dt = \int_0^x f'(t)dt = f(x),$$

qui montre bien que K est le noyau associé à $\mathcal{H}$.

Décompositions de RKHS

Propriété *Soit $\mathcal{H}$ un espace auto-reproduisant sur $\mathcal{T}$ dont le noyau K peut se décomposer en $K = K_0 + K_1$, avec K_0 et K_1 tous deux définis semi-positifs et tels que $K_0(x, \cdot) \in \mathcal{H}$ et $K_1(x, \cdot) \in \mathcal{H}$ pour tout $x \in \mathcal{T}$ et $\langle K_0(\cdot, x), K_1(\cdot, y) \rangle_{\mathcal{H}} = 0$, $\forall x, y \in \mathcal{T}$. Alors*

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1.$$

Réciproquement, si $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$ alors $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ a pour noyau reproduisant $K = K_0 + K_1$.

Preuve

Par définition de $\mathcal{H}$ et par orthogonalité de K_0 et K_1 on a

$$K_0(x,y) = \langle K_0(\cdot,x), K(\cdot,y)\rangle_{\mathcal{H}} = \langle K_0(\cdot,x), K_0(\cdot,y)\rangle_{\mathcal{H}},$$

et donc l $\mathcal{H}_0$ est un sous-espace fermé de $\mathcal{H}$. Soit $f \in \mathcal{H}$ et soit $f = f_0 + f_0^{\perp}$. On a :

$$f(x) = \langle K_0(\cdot,x), f\rangle_{\mathcal{H}} = f_0(x) + \langle K_1(\cdot,x), f_0^{\perp}\rangle_{\mathcal{H}}$$

ce qui montre que K_1 est le noyau reproduisant de $\mathcal{H} \ominus \mathcal{H}_0$. La réciproque est triviale.

Décomposition et FANOVA à un facteur

A titre d'exemple considérons un RKHS $\mathcal{H}$ sur $[0,1]$ de noyau K contenant les fonctions constantes et notons $\mathcal{H}_0$ l'espace vectoriel engendré par les fonctions constantes sur $[0,1]$ et $\mathcal{H}_1$ son supplémentaire orthogonal dans $\mathcal{H}$.

D'après ce qui précède chacun de ces sous-espaces est un RKHS avec pour noyaux respectifs $K_0(x,y) \equiv 1$ et $K_1(x,y) = K(x,y) - K_0(x,y)$, noyau du projecteur sur l'espace orthogonal aux constantes.

Ceci permet donc de définir une décomposition de la variance de $\mathcal{H}$, avec $\mathcal{H}_0$ l'espace des "moyennes" et $\mathcal{H}_1$ l'espace des "contrastes".

Les splines linéaires

Considérons l'espace de Sobolev $W^1([0,1])$ défini par

$$W^1([0,1]) = \left\{ f \in [0,1] \to \mathbb{R}, \text{ cont., dérivable p.p. , } f' \in L^2([0,1]) \right\},$$

et soient $K_0(x,y) = 1$ et $K_1(x,y) = \min(x,y)$. Il est facile de voir que

$\mathcal{H}_0 = \{f \in W^1; f' = 0\}$ $\mathcal{H}_1 = \{f \in W^1; f(0) = 0$ et $f' \in L^2\} = \mathcal{H}$ (exemple 1)

Ce noyau est utile pour le lissage par splines linéaires.

Plus généralement . . .

Soit $m \in \mathbb{N}_*$ et soit

$$\mathcal{H} = \left\{ f \in \mathbb{R}^{[0,1]},\ m-1 \text{ cont. dérivable}, f^{(m)} \in L^2, f^{(k)}(0) = 0, k = 0, m-1 \right\}.$$

C'est un espace de Hilbert lorsqu'on le muni du produit scalaire $\langle f, g \rangle_{\mathcal{H}} = \int_0^1 f^{(m)}(t) g^{(m)}(t) dt$ et la norme $\|f\|_{\mathcal{H}}$ mesure encore une fois la régularité de f. C'est un espace auto-reproduisant de noyau

$$K_1(x, y) = \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} \frac{(y-u)_+^{m-1}}{(m-1)!} du.$$

Splines polynomiales de lissage

Comme pour le cas des splines linéaires, soit $m \in \mathbb{N}_*$ et notons $W^m([0,1])$ l'espace de Sobolev sur $[0,1]$ d'ordre m de produit scalaire

$$\langle f, fg \rangle = \sum_{k=0}^{m-1} f^{(k)}(0) g^{(k)}(0) + \int_0^1 f^{(m)}(t) g^{(m)}(t) dt.$$

On a encore

$$W^m = \mathcal{H}_0 \oplus \mathcal{H}$$

avec

$$K(x,y) = \sum_{k=0}^{m-1} \frac{x^k}{k!} \frac{y^k}{k!} + K_1(x,y)$$

donnant pour $\mathcal{H}_0$ l'espace des polynômes de degré au plus $m-1$ et admettant le noyau $K_0(x,y) = \sum_{k=0}^{m-1} \frac{x^k}{k!} \frac{y^k}{k!}$.

Noyaux de Green et RKHS...

Soit $\mathcal{T} = \mathbb{R}^d$ et D un opérateur différentiel sur une classe de fonctions $\mathcal{H}$ sur $\mathcal{T}$ telle que munie du produit scalaire

$$\langle f, g \rangle_{\mathcal{H}} = \langle Df, Dg \rangle_{L^2(\mathcal{T})},$$

$\mathcal{H}$ soit un espace de Hilbert. Alors

Proposition. *$\mathcal{H}$ est un RKHS et son noyau reproduisant est la fonction de Green de l'opérateur D^*D où D^* est l'adjoint de D.*

Fonction de Green (rappels)

Considérons l'équation différentielle sur $\mathcal{H}$:

$$f = Dg,$$

avec g l'inconnue. Pour la résoudre on peut chercher g de la forme

$$g(\mathbf{x}) = \int_{\mathcal{T}} k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y},$$

pour une certaine fonction $k : \mathcal{T}^{\times 2} \to \mathbb{R}$ qui doit vérifier, pour tout $\mathbf{x} \in \mathcal{T}$,

$$f(\mathbf{x}) = Dg(\mathbf{x}) = \langle Dk_x, f \rangle_{L^2(\mathcal{T})}.$$

k est appelée fonction de Green de l'opérateur D.

Preuve de la proposition

Soient $\mathcal{H}$ un espace de Hilbert muni de

$$\langle f, g \rangle_{\mathcal{H}} = \langle Df, Dg \rangle_{L^2(\mathcal{T})}$$

et K la fonction de Green de l'opérateur D^*D. Pour tout $\mathbf{x} \in \mathcal{T}$ on a $K(\cdot, \mathbf{x}) = K_{\mathbf{x}}(\cdot) \in \mathcal{H}$ car

$$\langle DK_{\mathbf{x}}, DK_{\mathbf{x}} \rangle_{L^2(\mathcal{T})} = \langle D^*DK_{\mathbf{x}}, K_{\mathbf{x}} \rangle_{L^2(\mathcal{T})} = K(\mathbf{x}, \mathbf{x}) < \infty$$

D'autre part

$$f(\mathbf{x}) = \langle D^*DK_{\mathbf{x}}, f \rangle_{L^2(\mathcal{T})} = \langle DK_{\mathbf{x}}, Df \rangle_{L^2(\mathcal{T})} = \langle K_{\mathbf{x}}, f \rangle_{\mathcal{H}}.$$

Autre exemple: un espace de fonction analytiques

Soit $\mathbb{D} =]-1,1[$ le disque unité ouvert de $\mathbb{R}^d$ et soit $H^2(\mathbb{D})$ l'espace de Hardy des fonctions analytiques sur $\mathbb{D}$ dont les coefficients du développement en série entière sont de carré sommable, i.e.

$$H^2(\mathbb{D}) = \left\{ f \in \mathbb{R}^{\mathbb{D}}; f(x) = \sum_{k=0}^{\infty} a_k x^k \text{ avec } \sum_{k=0}^{\infty} a_k^2 < \infty \right\}$$

On définit sur $H^2(\mathbb{D})$ le produit scalaire $\langle f, g \rangle = \sum_{k=0}^{\infty} a_k b_k$ qui fait que $H^2(\mathbb{D})$ peut être identifié avec l'espace de Hilbert $\ell_2(\mathbb{N})$.

$H^2(\mathbb{D})$ est un RKHS

Soit $x \in \mathbb{D}$ et considérons $g_x(z) := \sum_{k=0}^{\infty} x^k z^k$. Clairement $g_x \in H^2(\mathbb{D})$ et

$$\langle f, g_x \rangle = \sum_{k=0}^{\infty} x^k a_k = f(x).$$

L'évaluation est donc continue et $H^2(\mathbb{D})$ est un RKHS ayant pour noyau, le noyau de Szego :

$$K(x, z) = \sum_{k=0}^{\infty} z^k x^k = \frac{1}{1 - xz}.$$

Autres espaces RKHS: domaines produits

Des sommes et des produits tensoriels d'espaces RKHS, ou des sous-espaces fermés de RKHS sont encore des RKHS, permettant ainsi la construction de RKHS dans des domaines assez généraux. Cela est principalement dû au fait que sommes et produits tensoriels de noyaux de type positif le sont encore.

Par exemple, en prenant $s_1, t_1 \in \mathcal{T}_1$, $s_2, t_2 \in \mathcal{T}_2$ et en notant $\mathbf{s} = (s_1, s_2)$ et $\mathbf{t} = (t_1, t_2)$, alors

$$K(\mathbf{s}, \mathbf{t}) = K_1(s_1, t_1) K_2(s_2, t_2)$$

est un noyau défini demi-positif sur $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2$ tant que K_1 et K_2 sont des noyau sur leurs domaines respectifs.

Produits tensoriels de RKHS

Si $\mathcal{H}_1$ est un RKHS sur $\mathcal{T}_1$ de noyau K_1 et $\mathcal{H}_2$ est un RKHS sur $\mathcal{T}_2$ de noyau K_2, d'après ce qui précède $K = K_1 \otimes K_2$ est un noyau défini positif sur $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2$ auquel correspond donc un RKHS, disons $\mathcal{H}$. Ce dernier n'est rien d'autre que le *produit tensoriel de* $\mathcal{H}_1$ *et de* $\mathcal{H}_2$, i.e. $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$.

Rappelons que si $f_1, f_2 \in \mathcal{H}_1$ et $g_1, g_2 \in \mathcal{H}_2$ alors

$$\langle f_1 \otimes g_1, f_2 \otimes g_2 \rangle_{\mathcal{H}} = \langle f_1, f_2 \rangle_{\mathcal{H}_1} \langle g_1, g_2 \rangle_{\mathcal{H}_2}.$$

Produits tensoriels de RKHS (suite)

Supposons que l'on dispose de RKHS $\mathcal{H}_\gamma$ sur des domaines $\mathcal{T}_\gamma$ de noyaux respectifs K_γ, $\gamma = 1, \ldots, \Gamma$ et supposons de plus que chaque espace $\mathcal{H}_\gamma$ admet une décomposition de type ANOVA, $\mathcal{H}_\gamma = \mathcal{H}_{0,\gamma} \oplus \mathcal{H}_{1,\gamma}$ où $\mathcal{H}_{0,\gamma} = \{f \in \mathcal{H}; f \propto 1\}$ et $K_{0,\gamma} \perp K_{1,\gamma}$. Alors l'espace tensoriel produit $\mathcal{H} = \otimes_{\gamma=1}^{\Gamma} \mathcal{H}_\gamma$ admet la décomposition

$$\mathcal{H} = \otimes_{\gamma=1}^{\Gamma} (\mathcal{H}_{0,\gamma} \oplus \mathcal{H}_{1,\gamma}) = \oplus_u \{(\otimes_{\gamma \in u} \mathcal{H}_{1,\gamma}) \otimes (_{\gamma \notin u} \mathcal{H}_{0,\gamma})\}$$

sur l'ensemble des parties u de $\{1, \ldots, \Gamma\}$.

Produits tensoriels de RKHS (suite)

On décompose ainsi une fonction de $\mathcal{H}$ définie sur $\mathcal{T}$ en "effets principaux" et "intéractions" selon, pour tout $\mathbf{x} \in \mathcal{T}$:

$$f(\mathbf{x}) = \sum_{u \subseteq \{1,2,\ldots,\Gamma\}} f_u(\mathbf{x}),$$

- f_u ne dépend que des composantes de $\mathbf{x}$ dont l'indice est dans u
- $\langle f_u, f_v \rangle_{\mathcal{H}} = 0$, $u \neq v$, $\quad f_\emptyset = \langle f, \otimes_{\gamma=1}^{\Gamma} K_{0,\gamma} \rangle_{\mathcal{H}}$ est la moyenne globale.

Nous reviendrons sur cet exemple lorsque nous traiterons de l'ANOVA par lissage spline.

RKHS et processus gaussiens

Nous avons vu que tout noyau défini positif sur un ensemble $\mathcal{T}$ permet de construire un espace RKHS. Or, tout processus gaussien centré indexé par $\mathcal{T}$ admet un noyau de covariance qui est un noyau de type positif ce qui permet d'établir la relation entre espace Gaussiens et RKHS (Neveu, Parzen, . . .).

On peut d'ailleurs pousser plus loin ces notions et ces constructions par l'étude des mesures Gaussiennes sur des espaces abstraits (Gross, Badrikian et Chevet, . . .).

Deuxième Partie

Lissage et RKHS

Le théorème du représentant

Théorème Soit $\mathcal{T}$ un ensemble muni d'un noyau d.p. K, $\mathcal{H}_K$ le RKHS associé, et $\mathcal{S} \subset \mathcal{T}$ un sous-ensemble de cardinal fini n. Soit $\mathbf{y} \in \mathbb{R}^n$ et soit $\Psi_{\mathbf{y}} : \mathbb{R}^{n+1} \to \mathbb{R}$ une fonction de $n+1$ arguments, croissante par rapport au dernier argument. Alors, toute solution au problème :

$$\min_{f \in \mathcal{H}_K} \Psi_{\mathbf{y}}(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n), \|f\|_{\mathcal{H}_K}) = \min_{f \in \mathcal{H}_K} \xi(f, \mathcal{S})$$

admet une représentation de la forme :

$$\forall \mathbf{x} \in \mathcal{T}, \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}).$$

Souvent, la fonction Ψ est de la forme ($\nu > 0$, c convexe) :

$$\Psi_{\mathbf{y},\nu}(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n), \|f\|_{\mathcal{H}_K}) = c(\mathbf{y}, f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) + \nu \|f\|_{\mathcal{H}_K}.$$

Preuve

Comme $\mathcal{H}_K$ est un espace de Hilbert auto-reproduisant il est strictement convexe et les évaluations sont continues. Pour monter le théorème, on remarquera d'abord que pour toute fonction $f \in \mathcal{H}_K$, il existe g dans un espace de dimension au plus n telle que

$$\Psi(g) \leq \Psi(f)$$

En effet, soit $F = \cap_i \ker(\delta_{\mathbf{x}_i})$. C'est un sous-espace fermé et si P_F désigne la projection sur F on a :

$$\|f - P_F(f)\|_{\mathcal{H}_K} \leq \|f\|_{\mathcal{H}_K} \quad \text{et} \quad (f - P_F(f))(\mathbf{x}_i) = f(\mathbf{x}_i),\ i = 1, \ldots, n.$$

Donc $\Psi(f - P_F(f)) \leq \Psi(f)$. Soit maintenant G l'e.v. engendré par les $K(\cdot, \mathbf{x}_i)$. Comme $G = F^{\perp}$, on a donc $f - P_F(f) = P_G(f)$ et c'est terminé.

Splines polynomiales de lissage

Considérons le problème de débruitage

$$Y_i = \eta(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

avec $\eta \in W^m([0,1])$ où $m \in \mathbb{N}_*$ et W_m est l'espace de Sobolev d'ordre m.

La spline de lissage polynomiale de degré m est la solution du problème variationnel

$$\min_{f \in W_m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \nu \int_0^1 (f^{(m)}(u))^2 du.$$

Splines de lissage (suite)

Nous remarquerons que la pénalité $J(f)$ est la norme de la projection de f sur l'espace orthogonal aux polynômes de degré au plus $m-1$ et est donc une semi-norme. Le théorème du représentant permet donc d'affirmer que la solution optimale est a rechercher parmi les fonctions $\eta \in \mathcal{H}_0 + \mathcal{H}_1$ de la forme

$$\eta(x) = \sum_{k=0}^{m-1} d_k \frac{x^k}{k!} + \sum_{i=1}^{n} c_i K_1(x_i, x),$$

avec $\mathbf{c} \in \mathbb{R}^n$ et $\mathbf{d} \in \mathbb{R}^m$ des vecteurs réels.

Splines de lissage (suite)

En notant S la matrice d'ordre $n \times m$ de (i,k)ième terme géneral $\frac{x_i^k}{k!}$ et Q la matrice de Gram associée au noyau K_1, i.e. $Q = (K_1(x_i, x_j))_{i,j=1,\ldots,n}$, les vecteurs optimaux $\mathbf{c}$ et $\mathbf{d}$ sont obtenus en minimisant

$$(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\nu\mathbf{c}^T Q\mathbf{c}.$$

Notant $M = Q + n\nu\mathbf{I}_n$, la solution est donnée par

$$\mathbf{d} = (S^T M^{-1} S)^{-1} S^T M^{-1}\mathbf{Y}, \quad \mathbf{c} = M^{-1}(\mathbf{I} - S(S^T M^{-1} S)^{-1} S^T M^{-1})\mathbf{Y}.$$

Evidemment les calculs ne se font pas avec ces formules.

Remarques

– On peut aussi dans le problème de débruitage considérer, plutôt que des évaluations des expressions du type $Y_i = L_i f + \epsilon_i,\ i = 1, \ldots, n,$ où les L_i sont des fonctionnelles bornées sur $\mathcal{H}_K$ et donc admettant un représentant $\phi_i \in \mathcal{H}_K$ plutôt que $\delta_{\mathbf{x}_i}$. On aborde ainsi des problèmes de régularisation de type Tikkhonov pour la résolution de pbs inverses.

– Lorsque le bruit est gaussien, le débruitage est un problème d'identification du max. de vrais. On peut donc remplacer la fonctionnelle à minimiser par des fonctionnelles telles que celles du théorème du représentant, lorsque l'on se pose des problèmes de régularisation de vraisemblance pénalisée.

Remarques (suite)

Dans l'expression de la fonctionnelle à minimiser dans le théorème du représentant, la fonctionnelle $c(\mathbf{y}, f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ est supposée strictement convexe de sorte à ce que le minimum soit unique.

Lorsque l'on traite des problèmes de vraisemblance pénalisée (par exemple débruitage de données binaires par régression logistique) cette fonction de perte est de la forme

$$c(y, f(\mathbf{x})) = -yf(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))).$$

Dans le cas de la classification de données binaires utilisée dans les SVM, la fonction de perte est

$$c(y, f(\mathbf{x})) = (1 - f(\mathbf{x})y)_+.$$

Choix des paramètres de régularisation pour splines

Soit $f_\nu^{[k]}(\cdot)$ le minimiseur de

$$\min_{f \in W_m} \frac{1}{n} \sum_{i=1, i \neq k}^{n} (Y_i - f(x_i))^2 + \nu \int_0^1 (f^{(m)}(u))^2 du.$$

L'estimateur (leaving-out-one) par validation croisée de ν est le minimiseur de

$$CV(\nu) = \frac{1}{n} \sum_{k=1}^{n} (Y_k - f_\nu^{[k]}(\mathbf{x}_k))^2.$$

Pour ν fixé, l'estimateur spline est une fonction linéaire des données, i.e. il existe une matrice de lissage $A(\nu)$ telle que $f_\nu(\mathbf{x}_i) = (A(\nu)\mathbf{Y})_i$.

Lemme

On a

$$CV(\nu) = \frac{1}{n}\sum_{k=1}^{n}\frac{(Y_k - f_\nu(\mathbf{x}_k))^2}{(1-a_{kk}(\nu))^2}$$

et pour des raisons d'invariance on estime plutôt ν en minimisant le critère de validation croisée géneralisée

$$GCV(\nu) = \frac{\frac{1}{n}\sum_{k=1}^{n}(Y_k - f_\nu(\mathbf{x}_k))^2}{(1-\frac{1}{n}\sum_{\ell=1}^{n}a_{\ell,\ell}(\nu))^2}$$

Dans le cas où la variance du bruit est connue on minimise un estimateur du risque sans biais

$$R(\nu) = \|(I - A(\nu))\mathbf{Y}\|^2 + 2\sigma^2 \mathsf{trace}(A(\nu)).$$

Analyse de la variance fonctionnelle (SS-ANOVA)

Concept initial (Antoniadis (1983)) - ensuite SS-ANOVA (Wahba (1990))

Ces modèles utilisant des RKHS couvrent un aspect très général pour l'analyse de données.

- ANOVA fonctionnelle
- Décompositions FANOVA
- RKHS pour décompositions FANOVA
- Exemples

Analyse de la variance fonctionnelle

Le modèle gaussien général s'écrit sous la forme

$$y_i = f(t_1(i), \cdots, t_d(i)) + \epsilon_i, \quad i = 1, \cdots, n,$$

où

- $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(0, \sigma^2 I_{n\times n})$,
- $t_\alpha \in \mathcal{T}^{(\alpha)}$, où $\mathcal{T}^{(\alpha)}$, $\alpha = 1, \cdots, d$ sont des ensembles mesurables donnés.
- $\mathcal{T} = \mathcal{T}^{(1)} \times \cdots \times \mathcal{T}^{(d)}$, et
- σ^2 est éventuellement inconnu.

Pour f satisfaisant à certaines conditions de mesurabilité une décomposition unique de type ANOVA

$$f(t_1, \cdots, t_d) = \mu + \sum_{\alpha} f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_{\alpha\beta}) + \cdots$$

peut être obtenue comme suit:

Soit $d\mu_\alpha$ une probabilité sur $\mathcal{T}^{(\alpha)}$ et considérons l'opérateur de moyenne $\mathcal{E}_\alpha$ sur $\mathcal{T}$ définit par

$$(\mathcal{E}_\alpha f)(t) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \cdots, t_d) d\mu_\alpha(t_\alpha).$$

L'identité est alors décomposée comme

$$I = \prod_{\alpha}(\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)) = \prod_{\alpha}\mathcal{E}_\alpha + \sum_{\alpha}(I - \mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta$$

$$+ \sum_{\alpha<\beta}(I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma + \cdots + \prod_{\alpha}(I - \mathcal{E}_\alpha).$$

Les composantes de cette décomposition génèrent la décomposition ANOVA de f avec

$$\mu = (\prod_{\alpha}\mathcal{E}_\alpha)f, \quad f_\alpha = ((I - \mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta)f,$$

$$f_{\alpha\beta} = ((I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma)f$$

et ainsi de suite.

L'idée est de construire un RKHS $\mathcal{H}$ de fonctions sur $\mathcal{T}$ de sorte que les composantes de la décomposition ANOVA constituent une décomposition orthogonale de f dans l'espace $\mathcal{H}$. Plus précisément soit $\mathcal{H}^{(\alpha)}$ un RKHS de fonctions sur $\mathcal{T}^{(\alpha)}$ avec $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha) d\mu_\alpha = 0$ pour $f_\alpha(t_\alpha) \in \mathcal{H}^{(\alpha)}$, et soit $[1^{(\alpha)}]$ l'e.v. des fonctions constantes sur $\mathcal{T}^{(\alpha)}$. Par abus de notation, on identifiera par la suite $\mathcal{H}^{(\alpha)}$ au sous-espace de $\mathcal{H}$ suivant $[1^{(1)}] \otimes \cdots \otimes [1^{(\alpha-1)}] \otimes \mathcal{H}^{(\alpha)} \otimes [1^{(\alpha+1)}] \otimes \cdots \otimes [1^{(d)}]$. On construit $\mathcal{H}$ selon

$$\mathcal{H} = \prod_{\alpha=1}^{d} (\{[1^{(\alpha)}]\} \oplus \{\mathcal{H}^{(\alpha)}\}) = [1] \oplus \sum_{\alpha} \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots .$$

Les composantes de cette décomposition ANOVA sont dans des sous-espaces de $\mathcal{H}$ mutuellement orthogonaux. Bien évidement cela dépend des mesures $d\mu_\alpha$ qui sont choisies en fonction de l'application visée.

SS-ANOVA (suite)

Ensuite, chaque $\mathcal{H}^{(\alpha)}$ peut éventuellement être décomposé en une partie paramétrique et une partie lisse, selon $\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$, où $\mathcal{H}_\pi^{(\alpha)}$ est de dimension finie (la partie "paramétrique") et $\mathcal{H}_s^{(\alpha)}$ (la partie "lisse") est le sup. orthogonal de $\mathcal{H}_\pi^{(\alpha)}$ dans $\mathcal{H}^{(\alpha)}$.

Cette écriture à l'aide de produits tensoriels de RKHS est alors très utile dans toute procédure de régularisation. On peut regrouper par exemple en un sous espace $\mathcal{H}^0$ les parties que l'on ne désire pas pénaliser et décomposer $\mathcal{H} = \mathcal{H}^0 \oplus \sum_\beta \mathcal{H}^\beta$.

SS-ANOVA (suite)

Supposons avoir choisi un modèle de régularisation $\mathcal{M}$, i.e. que l'on a regroupé en $\mathcal{H}^0$, de dimension M, les sous-espaces que nous ne pénaliserons pas et que nous avons renommé les autres espaces $\mathcal{H}^\beta, \beta = 1, 2, \cdots, p$.

En posant $\mathcal{H} = \mathcal{H}^0 \oplus \sum_\beta \mathcal{H}^\beta$, le problème d'estimation pénalisée devient alors : Déterminer f dans $\mathcal{H} = \mathcal{H}^0 \oplus \sum_\beta \mathcal{H}^\beta$ qui minimise

$$\Psi_\lambda(\mathbf{Y}, f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(t(i)))^2 + \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} \|P^\beta f\|^2,$$

où P^β est le projecteur orthogonal de $\mathcal{H}$ sur $\mathcal{H}^\beta$, et choisir les paramètres de régularisation λ, θ_β pour réaliser le meilleur compromis biais-variance.

SS-ANOVA (suite)

Si $\{\phi_1, \ldots, \phi_M\}$ est une base de $\mathcal{H}^0$, l'espace $\mathcal{H}_1 = \sum_\beta \mathcal{H}^\beta$ est un sous espace fermé de $\mathcal{H}$ et est donc un RKHS dont le noyau K_1 sur $\mathcal{T}$ peut s'écrire sous la forme

$$K_1(\mathbf{x}, \mathbf{y}) = \sum_{\beta=1}^{p} \theta_\beta^{-1} K_\beta(\mathbf{x}, \mathbf{y}).$$

Notons que $K_\beta(\mathbf{x}, \mathbf{y})$ ne dépendent que d'un sous-ensemble des composantes de $(\mathbf{x}, \mathbf{y})$. Les pondérations θ_β^{-1} permettent une régularisation différente pour chaque composante. D'après le théorème du représentant la fonction f_λ minimisant $\Psi_\lambda(\mathbf{Y}, f)$ s'écrit donc

$$f_\lambda(\cdot) = \sum_{k=1}^{M} d_k \phi_k(\cdot) + \sum_{i=1}^{n} c_i \sum_{\beta=1}^{p} \theta_\beta^{-1} K_\beta(\mathbf{t}_i, \cdot).$$

Un exemple d'analyse

Nous nous proposons d'analyser les données d'ozone d'Arosa (Andrews et Hezberg (1985)). Il s'agit des moyennes mensuelles d'enregistrements de l'épaisseur de la couche d'ozone (en unités Dobson) au dessus d' Arosa (Suisse) de 1926 à 1971. Nous nous intéresserons aux effets de la période et de l'année sur l'épaisseur.

Nous sommes ici en présence de deux facteurs ($d = 2$), le mois et l'année, que nous allons traiter pour l'illustration, comme continus, i.e. $\mathcal{T}_1 = [1, 12]$ et $\mathcal{T}_2 = [1926, 1971]$. Pour des raisons pratiques, nous allons transformer les donnés de sorte que chaque facteur varie dans $[0, 1]$.

L'effet du mois sera modélisé par des splines périodiques d'ordre 2 sur $[0, 1]$ alors que celui des années par des splines cubiques sur $[0, 1]$.

Splines périodiques (Wahba(1990))

Pour les fonctions splines périodiques d'ordre m sur $\mathcal{T} = [0,1]$ on a

$$\mathcal{H} = \{f : f^{(j)} \text{ abs. cont.}, f^{(j)}(0) = f^{(j)}(1), j = 0, \ldots, m-1, \int_0^1 (f^{(m)}(t))^2 dt < \infty\}$$

On note souvent $\mathcal{H} = W_m(per)$. On a

$$\mathcal{H}_0 = \text{span}\{\mathbf{1}\}, \quad K_1(s,t) = \sum_{k=1}^{\infty} \frac{2}{(2\pi k)^{2m}} \cos 2\pi k(s-t),$$

$$\|P_1 f\|^2_{\mathcal{H}} = \int_0^1 (f^{(m)}(t))^2 dt$$

Ozone (suite)

Rappelons que l'espace des splines cubiques sur $[0,1]$ se décompose en

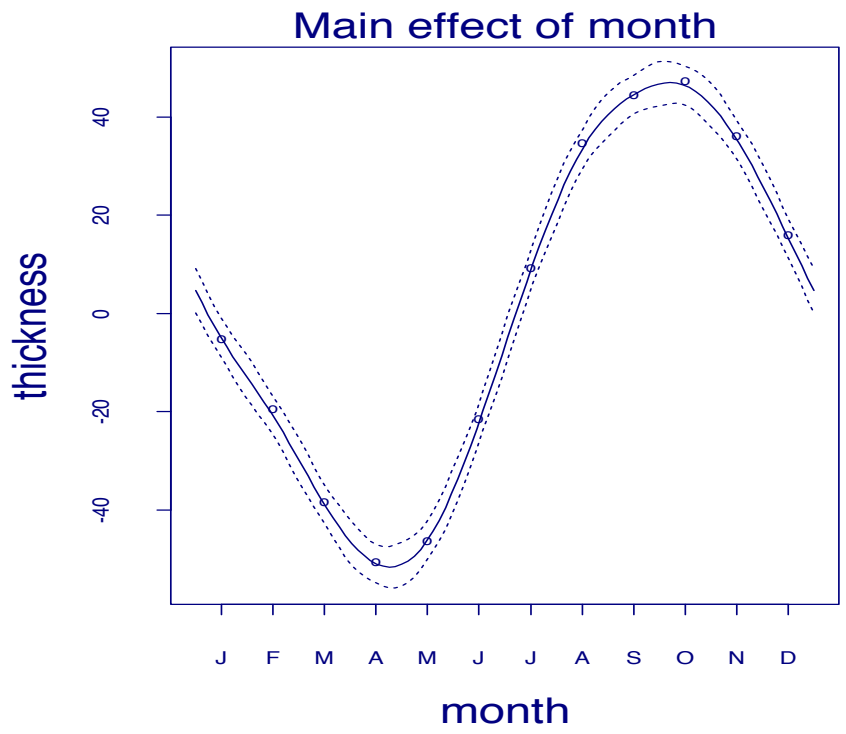
$$\mathcal{H} = \{1\} \oplus \{t\} \oplus \left\{ g \in W^2([0,1]) : \int g(t)dt = \int g'(t)dt = 0 \right\}.$$

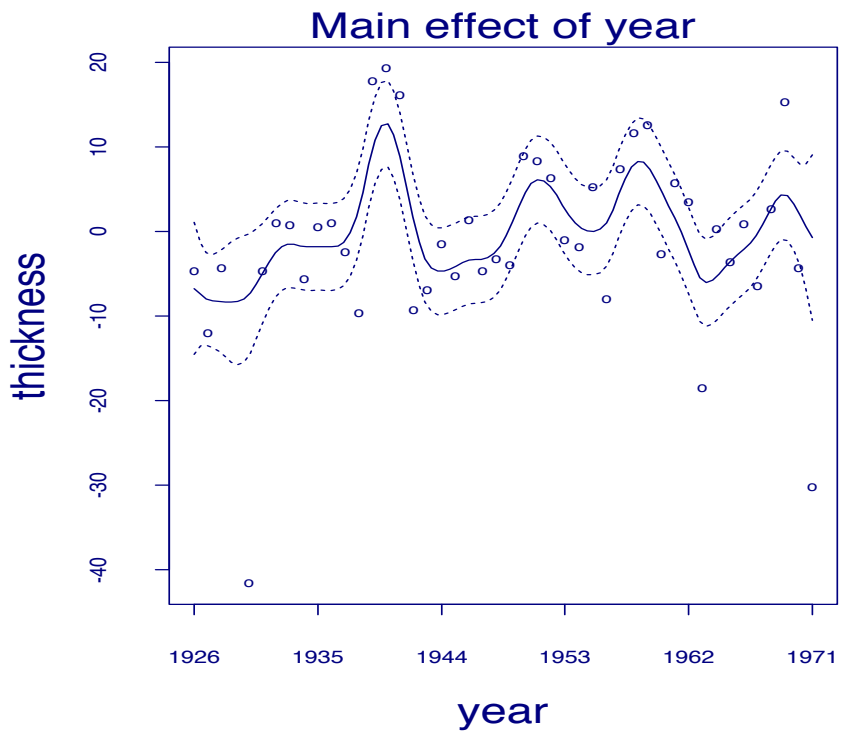
En considérant donc cette décomposition pour le facteur "année" (disons t_2) et la décomposition périodique $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1$ pour le facteur "mois" (disons t_1), cela revient à adopter le modèle suivant pour l'épaisseur moyenne f :

$$f(t_1, t_2) = \mu + \beta t_2 + f_1(t_1) + f_2(t_2) + g_1(t_1)t_2 + f_{1,2}(t_1, t_2).$$

Commandes R

```
 > data(Arosa)
> Arosa$csmonth <- (Arosa$month-0.5)/12
> attach(Arosa)
> csyear <- (year-1)/45
> ozone.fit <- ssr(thick~I(csyear-0.5), spar="m",
+ rk=list(periodic(csmonth), cubic(csyear),
+ rk.prod(periodic(csmonth), kron(csyear-.5)),
+ rk.prod(periodic(csmonth), cubic(csyear))))
> summary(ozone.fit)
      Coefficients (d):
        (Intercept) I(csyear - 0.5)
         336.82981         6.03153
     GML estimate(s) of smoothing parameter(s) :
      1.531358e-06 2.779106e-07 6.174975e-01 2.026814e-02
Estimate of sigma:  15.84154
```


Main effect of month
thickness
40
20
0
-20
-40
J
F
M
A
M
J
J
A
S
O
N
D
month


Main effect of year
thickness
20
10
0
-10
-20
-30
-40
1926
1935
1944
1953
1962
1971
year

Références

Livres

Hastie, T., Tibshirani, R. & Friedman, J. (2001), The Elements of Statistical Learning; Data mining, Inference and Prediction, Springer Verlag, New York.
F. Riesz and B. Sz. Nagy. *Functional Analysis*. Ungar, New York, 1955
G. Wahba. *Spline Models for Observational Data*. SIAM, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics

Papiers

N. Aronszan. Theory of reproducing kernels. Transactions of the American Mathematical Society, 68:337–404, 1950.
E. Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt, editor, Proceedings of the Symposium on Time Series Analysis, pages 155–169. Wiley, 1963.
G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smooth- ing problem. Ann. Statist., 13:1378–1402, 1985
C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. J. Royal Statistical Soc. Ser. B, 55:353–368, 1993.