

Processus stochastiques pour la classification selon le style postural

Christophe Denis
Université Paris Descartes
MAP5, UMR CNRS 8145

En collaboration avec
A. Chambaz (MAP5, UMR CNRS 8145)
A. Samson (MAP5, UMR CNRS 8145)
et P-P. Vidal (CESEM, UMR CNRS 8194)

17/11/2011
Télécom ParisTech

Introduction

Modélisation

Procédure de classification

Application aux données réelles

Perspectives

Annexe

Introduction

Modélisation

Procédure de classification

Application aux données réelles

Perspectives

Annexe

Contexte

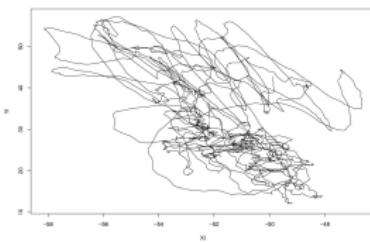
- Le maintien postural est le fruit du traitement dynamique, par le cerveau, d'informations provenant des systèmes :
 - ▶ visuel,
 - ▶ vestibulaire,
 - ▶ proprioceptif.
- Une stratégie optimale de maintien postural repose sur l'exploitation des trois types d'informations.
- A contrario, privilégier un unique type d'information (typiquement visuelle) est une stratégie ne permettant pas de gérer l'imprévu.
- Les objectifs de l'étude du maintien postural sont de permettre :
 - ▶ à un médecin généraliste d'identifier chez un patient, de manière simple, un trouble de la posture ;
 - ▶ la mise au point de programme de physiothérapie adapté au patient.

Notre jeu de données

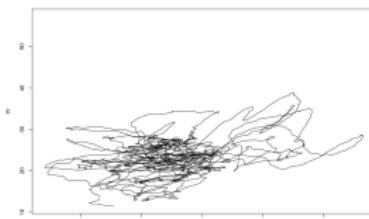
- 54 patients : 32 patients sains, 22 hémiplégiques.
- Pour chaque patient :
 - ▶ covariables W_t (âge, genre, poids, taille, latéralité)
 - ▶ mesures $(X_t)_{t \in T}$ prises lors de différents protocoles.
- Pour chaque protocole, le centre de pression maximale exercée de chaque pied a été enregistré au cours du temps sur une plateforme de force :
 - ▶ $X_t = (L_t, R_t)$ où $L_t = (L_t^1, L_t^2) \in \mathbb{R}^2$ et $R_t = (R_t^1, R_t^2) \in \mathbb{R}^2$.
- Quatre protocoles pris en compte :

protocole	0→15s	15→50s	50→70s
1		yeux fermés	
2		stimulation musculaire	
3		yeux fermés stimulation musculaire	
4		stimulation optokinétique	

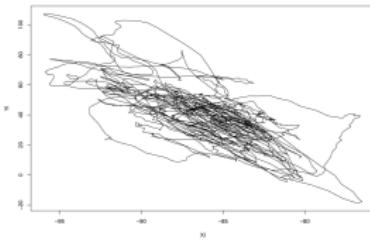
Exemple de 2 trajectoire $X_t = (L_t, R_t)$



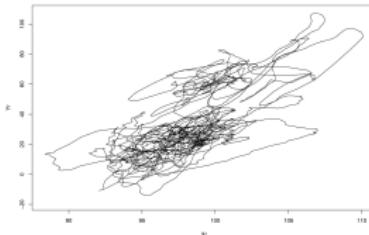
pied gauche (sujet sain)



pied droit (sujet sain)



pied gauche (sujet hémiplégique)



pied droit (sujet hémiplégique)

Objectifs principaux de l'étude

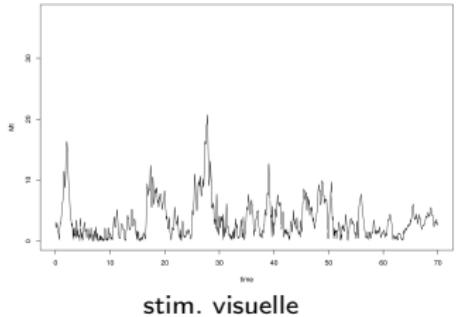
- Classer les patients en terme de maintien postural.
- Limiter le nombre de protocoles nécessaires pour procéder à la classification ;
 - ▶ classement des protocoles du plus au moins informatif en termes de maintien postural

Transformation des données brutes

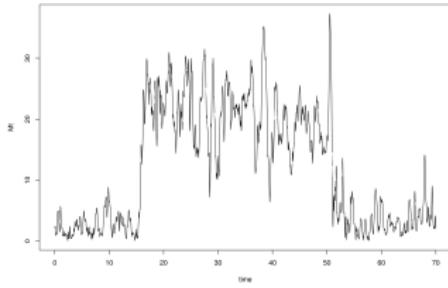
- Les mesures $(X_t)_{t \in T}$ sont de très grande dimension.
- On propose de substituer à $(X_t)_{t \in T}$ une quantité Y de faible dimension.
- Première étape : réduire $(X_t)_{t \in T}$ à $(C_t)_{t \in T}$, avec $C_t \in \mathbb{R}$
 1. pour tout $t \in T$, soit $B_t = (B_t^1, B_t^2)$
 - B_t^1 barycentre de (L_t^1, R_t^1)
 - B_t^2 barycentre de (L_t^2, R_t^2)
 2. choix d'un point de référence b ,
 3. définition de

$$C_t = \|B_t - b\|_2.$$

Trajectoire $t \mapsto C_t$



stim. visuelle

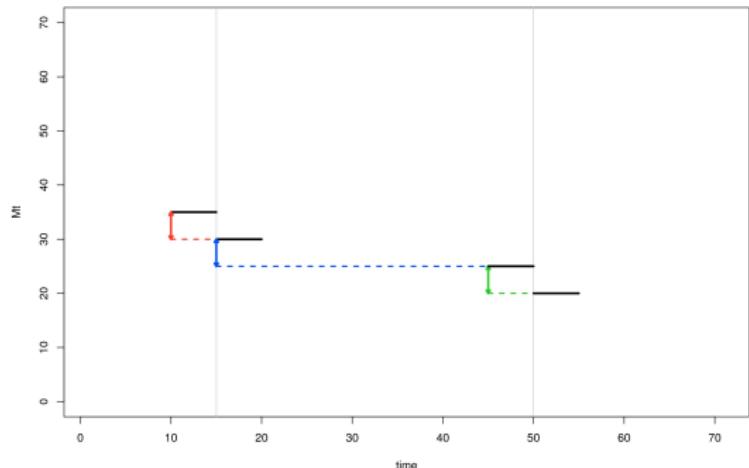


stim. musculaire et visuelle

Première approche : construction des mesures résumées

- On commence par considérer :
 - ▶ \bar{C}_1^- moyenne de C_t sur les 5 secondes précédant le début de la perturbation
 - ▶ \bar{C}_1^+ moyenne de C_t sur les 5 secondes succédant au début de la perturbation
 - ▶ \bar{C}_2^- moyenne de C_t sur les 5 secondes précédant la fin de la perturbation
 - ▶ \bar{C}_2^+ moyenne de C_t sur les 5 secondes succédant à la fin de la perturbation
- Puis on définit les mesures résumées $(Y_i)_{i \in \{1,2,3\}}$:

- ▶ $Y_1 = \bar{C}_1^+ - \bar{C}_1^-$
- ▶ $Y_2 = \bar{C}_2^- - \bar{C}_1^+$
- ▶ $Y_3 = \bar{C}_2^+ - \bar{C}_2^-$



Première approche : méthode de classification

- Observation $O = (W, A, Y^1, Y^2, Y^3, Y^4)$ de loi $P_0 \in \mathcal{M}$ où :
 - ▶ $W \in \mathbb{R} \times \{0, 1\}^2 \times \mathbb{R}^2$ vecteur des covariables,
 - ▶ $A \in \{0, 1\}$ indique l'état du patient (sain ou hémiplégique),
 - ▶ $Y^j \in \mathbb{R}^3$ est la mesure résumée associée au j -ème protocole.
- Procédure de classification en deux étapes :
 - ▶ Classement des protocoles selon leur pertinence (défini rigoureusement via un paramètre statistique)
 - ▶ Construction de quatre classifieurs $\phi^1, \phi^2, \phi^3, \phi^4$ fondés respectivement sur le meilleur protocole, les deux meilleurs, les trois meilleurs et les quatre protocoles.

Première approche : application aux données réelles

- Classement des protocoles : deux protocoles jugés très informatifs
 - ▶ protocole 1 : stimulation musculaire
 - ▶ protocole 2 : stimulation musculaire et visuelle
- Evaluation des performances de classifications par leave-one-out :

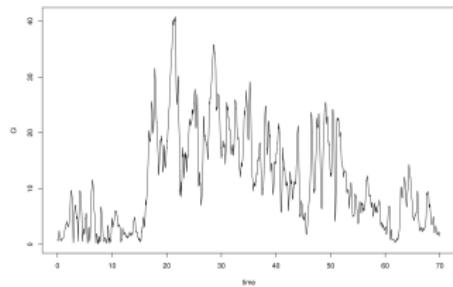
	ϕ^1	ϕ^2	ϕ^3	ϕ^4
% de bien classé	74%	81%	78%	85%

Première approche : avantages/inconvénients

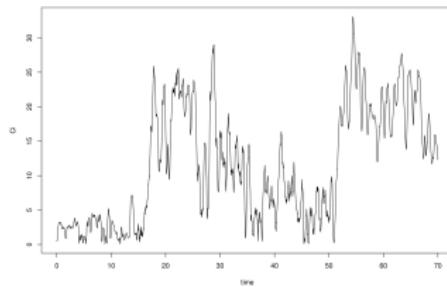
- Avantages :
 - ▶ procédure de classification facilement généralisable
 - ▶ amélioration possible des performances de classification avec un autre choix de mesures résumées.
- Inconvénients (*i.e.* pistes pour améliorer) :
 - ▶ mesures résumées proposées ne prennent pas en compte la dynamique temporelle des données
 - ▶ les ruptures provoquées par la phase de perturbation n'ont pas nécessairement lieu à 15 et 50 secondes exactement.
 - ▶ le pas de 5 secondes pour la construction des mesures résumées est arbitraire.
- Objectif de la deuxième approche :
 - ▶ Construction de nouvelles mesures résumées basées sur la dynamique de $(C_t)_{t \in \mathcal{T}}$.
 - ▶ Classification basée sur les deux protocoles les plus informatifs.

Idée

- C_t observé au temps $i\delta$, $\delta = 0.025s$ et $i \in \{1, \dots, N = 2800\}$.
- Notation : $C_{i:j} = (C_i, \dots, C_j)$.



stim. musculaire et visuelle



stim. musculaire

Introduction

Modélisation

Procédure de classification

Application aux données réelles

Perspectives

Annexe

Modélisation de $C_{1:N}$ (1/3)

- Trajectoire $C_{1:N}$ considérée comme l'observation d'un processus stochastique $(C(t))_{[T_0, T]}$ défini par :

$$\begin{cases} dC(t) = b(C(t), \phi)dt + a(C(t), \sigma)dW(t) \\ C(T_0) = C_0, \end{cases}$$

- ▶ $W(t)$ processus de Wiener standard,
- ▶ b le drift et a la volatilité sont des fonctions paramétriques connues.
- ▶ ϕ et σ paramètres inconnus.
- ▶ $T_0 = 0.025s$ et $T = 70s$.

Modélisation de $C_{1:N}$ (2/3)

- Phase de perturbation d'un protocole implique le changement des paramètres de $(C(t))_{t \in [\tau_0, T]}$:
 - ▶ au voisinage du début et de la fin de la phase de perturbation
 - ▶ éventuellement aussi pendant.
- On note la suite des instants de ruptures,

$$T_0 = \delta < T_1 = \tau_1 \delta < \dots < T_{k-1} = \tau_{k-1} \delta < T_k = T$$

- ▶ avec $K = 3$ ou 4 correspondant à 2 ou 3 instants de ruptures.
- Sur $[T_{k-1}, T_k]$:

$$\begin{cases} dC(t) = b(C(t), \phi_k)dt + a(C(t), \sigma_k)dW(t) \\ C(T_{k-1}) = C_{\tau_{k-1}}, \end{cases}$$

- Dernière étape : modélisation des formes paramétrique a et b .

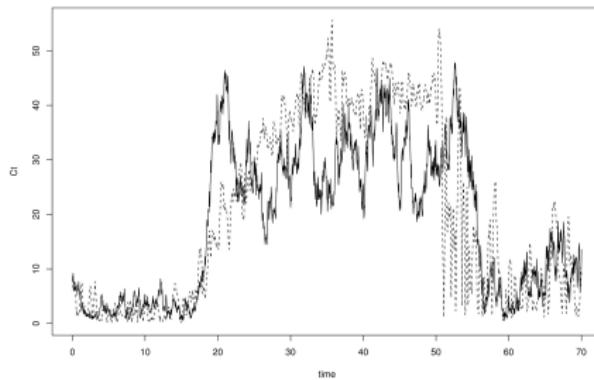
Modélisation de $C_{1:N}$ (3/3)

- On modélise $C(t)$ par un processus de Cox-Ingersoll-Ross (CIR).
- Modèle : pour tout $k \in \{1, \dots, K\}$, $(C(t))_{t \in [T_{k-1}, T_k]}$ est définie par :

$$\begin{cases} dC(t) = \lambda_k(\mu_k - C(t))dt + \sigma_k \sqrt{C(t)}dW(t) \\ C(T_{k-1}) = C_{\tau_{k-1}}, \end{cases}$$

- ▶ avec $C_{\tau_{k-1}}$ observation au temps T_{k-1} ,
- ▶ λ_k est l'échelle de temps du processus,
- ▶ μ_k est la moyenne de la distribution asymptotique,
- ▶ σ_k^2 est le paramètre de variance du processus.
- Propriétés élémentaires de $(C(t))_{t \in [T_{k-1}, T_k]}$:
 - ▶ Si $\frac{2\mu_k \lambda_k}{\sigma_k^2} \geq 1$ alors le processus reste positif et admet une distribution stationnaire,
 - ▶ loi Gamma de paramètres $\frac{2\mu_k \lambda_k}{\sigma_k^2}$ et $\frac{\sigma_k^2}{2\lambda_k}$.

Exemple de trajectoire



Trajectoire $C_{1:N}$ (trait plein)

Trajectoire simulée (pointillés)

Procédure d'estimation des paramètres du processus $C(t)$.

- On note $\theta_k = (\lambda_k, \mu_k, \sigma_k)$.
- Estimation des paramètres en deux étapes :
 - ▶ estimation de la suite des instants de ruptures $\Rightarrow \{\hat{T}_1, \dots, \hat{T}_{K-1}\}$.
 - ▶ pour tout $k \in \{1, \dots, K\}$, estimation du paramètre θ_k de $(C(t))_{[\hat{T}_{k-1}, \hat{T}_k]}$.
- Estimation de la suites des instants de ruptures fondée sur la minimisation d'un contraste empirique [Lavielle (2005)].
- Estimation des paramètres du processus fondée sur la minimisation d'un contraste empirique [Kessler (1997)].

Estimation des instants de rupture

- Cadre des modèles constants par morceaux.
- Contraste proposé par Lavielle (2005) :

$$\mathcal{J}(\tau, C_{1:N}) = \frac{1}{N} \sum_{k=1}^K (\tau_k - \tau_{k-1} + 1) \log(\gamma_{\tau_{k-1}:\tau_k}^2),$$

- ▶ où $\gamma_{\tau_{k-1}:\tau_k}^2$ estimateur empirique de la variance de $C_{\tau_{k-1}:\tau_k}$.
- l'estimateur $\hat{\tau}_N$ de la suite des instants de rupture est définie par :

$$\hat{\tau}_N = \arg \min_{\tau} \mathcal{J}(\tau, C_{1:N}).$$

- Sous des hypothèses générales, pour tout $k \in \{1, \dots, K-1\}$:

$$P(|\hat{\tau}_{N,k} - \tau_k| > \eta) \rightarrow 0, \text{ quand } \eta \rightarrow \infty \text{ et } N \rightarrow \infty.$$

Estimation des paramètres du processus

- Pour tout $k \in \{1, \dots, K-1\}$, l'approximation discrète par le schéma d'Euler Marumaya de $(C(t))_{[\hat{\tau}_{k-1}, \hat{\tau}_k]}$ est définie, $i \in \{\hat{\tau}_{k-1}, \dots, \hat{\tau}_k\}$ par :

$$\tilde{C}_{i+1} = (1 - \delta \lambda_k) \tilde{C}_i + \delta \lambda_k \mu_k + \sigma_k \sqrt{\delta} \sqrt{\tilde{C}_i} \eta_{i+1},$$

- Le contraste $\mathcal{L}(\theta_k)$ proposé par Kessler (1997) :

$$\mathcal{L}_k(\theta_k) = \sum_{i=\hat{\tau}_{k-1}}^{\hat{\tau}_k-1} \frac{(C_{i+1} - \theta_{1,k} C_i - \theta_{2,k})^2}{C_i \theta_{3,k}^2} + (\hat{\tau}_k - \hat{\tau}_{k-1} + 1) \log(\theta_{3,k}^2),$$

avec $\theta_{1,k} = 1 - \lambda_k \delta$, $\theta_{2,k} = \delta \lambda_k \mu_k$ et $\theta_{3,k}^2 = \delta \sigma_k^2$.

- On a $\hat{\theta}_k = \arg \min_{\theta_k \in \mathbb{R}_+^3} \mathcal{L}_k(\theta_k)$
- Sous des hypothèses générales on peut montrer la consistante des estimateurs $\hat{\theta}_k$.

Estimation des instants de ruptures sur les données réelles

Sujets normaux

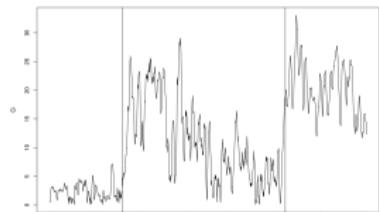
protocole	$K = 3$		$K = 4$		
1	18.28 (4.37)	50.66 (10.53)	16.62 (5.02)	31.90 (13.75)	52.97 (6.85)
2	19.75 (7.83)	52.62 (4.87)	15.84 (4.32)	31.59 (12.47)	52.68 (4.80)

Sujets hémiplégiques

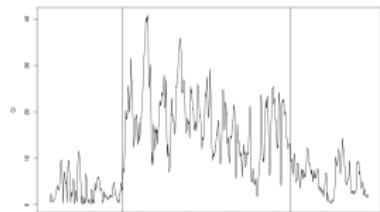
protocole	$K = 3$		$K = 4$		
1	15.27 (3.13)	45.27 (13.78)	14.27 (3.89)	31.18 (11.13)	51.86 (8.84)
2	16.95 (5.39)	48.13 (11.60)	13.95 (7.15)	30.45 (14.38)	55.36 (3.35)

Illustration de l'estimation des temps de ruptures

$K = 3$

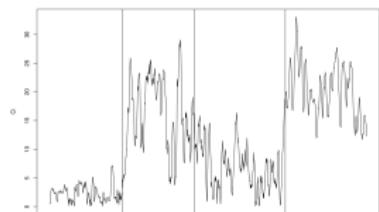


protocole 1

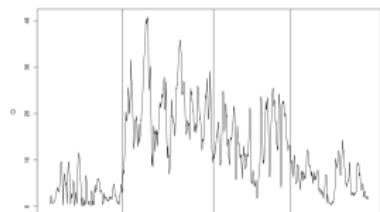


protocole 2

$K = 4$



protocole 1



protocole 2

Estimation des paramètres du processus sur les données réelles

protocole	param.	Sujets normaux		
		$k = 1$	$k = 2$	$k = 3$
1	$\theta_{1,k}$	0.98 (0.01)	0.99 (0.01)	0.98 (0.02)
	$\theta_{2,k}$	0.06 (0.04)	0.14 (0.12)	0.11 (0.08)
	$\theta_{3,k}^2$	0.04 (0.03)	0.03 (0.04)	0.03 (0.04)
2	$\theta_{1,k}$	0.98 (0.01)	0.99 (0.01)	0.98 (0.02)
	$\theta_{2,k}$	0.09 (0.08)	0.22 (0.26)	0.15 (0.14)
	$\theta_{3,k}^2$	0.06 (0.06)	0.05 (0.04)	0.05 (0.06)

protocole	param.	Sujets hémiplégiques		
		$k = 1$	$k = 2$	$k = 3$
1	$\theta_{1,k}$	0.96 (0.02)	0.98 (0.01)	0.98 (0.02)
	$\theta_{2,k}$	0.13 (0.09)	0.27 (0.30)	0.16 (0.13)
	$\theta_{3,k}^2$	0.08 (0.05)	0.05 (0.04)	0.05 (0.05)
2	$\theta_{1,k}$	0.96 (0.02)	0.98 (0.02)	0.97 (0.02)
	$\theta_{2,k}$	0.15 (0.10)	0.37 (0.36)	0.36 (0.62)
	$\theta_{3,k}^2$	0.1 (0.06)	0.1 (0.10)	0.09 (0.08)

Introduction

Modélisation

Procédure de classification

Application aux données réelles

Perspectives

Annexe

Modèle statistique pour la classification

- Construction des nouvelles mesures résumées fondée sur l'estimation des temps de rupture et des paramètres du processus.
- Plus formellement, à une trajectoire $C_{1:N}$ on associe la mesure résumée Y_K définie pour $K \in \{3, 4\}$ par :

$$Y_K = \left(\hat{\mu}_1, \dots, \hat{\mu}_K, \hat{\sigma}_1^2, \dots, \hat{\sigma}_K^2, \hat{T}_1, \dots, \hat{T}_{K-1} \right),$$

- On note Y_K^j la mesure résumée associée au protocole j .
- Observation :

$$O = (W, Y_3^1, Y_4^1, Y_3^2, Y_4^2, A) \sim P_0.$$

Définition des classifieurs

- On note $Z = (W, Y_3^1, Y_4^1, Y_3^2, Y_4^2)$ et \mathcal{S} l'ensemble des classifieurs fonction de Z .
- Pour tout $j \in \{1, 2\}$, $K \in \{3, 4\}$ on définit $S_K^j \in \mathcal{S}$:

$$S_K^j(W, Y_K^j) = \mathbb{1}\{\eta_K^j(W, Y_K^j) \geq \frac{1}{2}\},$$

- ▶ avec $\eta_K^j(W, Y_K^j) = P_0(A = 1 | W, Y_K^j)$;
- ▶ le classifieur S_K^j est le classifieur de Bayes basé sur W et Y_K^j .
- Pour tout $(K_1, K_2) \in \{3, 4\}$, on définit $S_{K_1, K_2} \in \mathcal{S}$:

$$S_{K_1, K_2}(W, Y_{K_1}^1, Y_{K_2}^2) = \mathbb{1}\{\eta_{K_1, K_2}(W, Y_{K_1}^1, Y_{K_2}^2) \geq \frac{1}{2}\},$$

- ▶ $\eta_{K_1, K_2}(W, Y_{K_1}^1, Y_{K_2}^2) = P_0(A = 1 | W, Y_{K_1}^1, Y_{K_2}^2)$
- ▶ le classifieur S_{K_1, K_2} est le classifieur de Bayes basé sur W et $(Y_{K_1}^1, Y_{K_2}^2)$.

Procédure de classification

- On définit pour $s \in S$, $R^{(P_0)}(s) = P_0(s(Z) \neq A)$.
- On définit les classifieurs ϕ_1 et ϕ_2 comme :

$$\phi_1 \in \arg \min_{S_K^j} R^{(P_0)}(S_K^j)$$

$$\phi_2 \in \arg \min_{S_{K_1, K_2}} R^{(P_0)}(S_{K_1, K_2}).$$

- Objectif : construire $\hat{\phi}_1$ et $\hat{\phi}_2$ estimateur de ϕ_1 et ϕ_2 .
- Procédure de classification en deux étapes :
 - ▶ construction des ensembles \mathcal{H}_1 et \mathcal{H}_2 d'estimateurs candidats pour ϕ_1 et ϕ_2 :
 - $\mathcal{H}_1 = \{\hat{S}_3^1, \hat{S}_4^1, \hat{S}_3^2, \hat{S}_4^2\}$
 - $\mathcal{H}_2 = \{\hat{S}_{3,3}, \hat{S}_{3,4}, \hat{S}_{4,3}, \hat{S}_{4,4}\}$
 - ▶ sélection de $\hat{\phi}_1$ (resp. $\hat{\phi}_2$) parmi l'ensemble \mathcal{H}_1 (resp. \mathcal{H}_2) fondée sur le principe de validation croisée.

Construction de l'ensemble \mathcal{H}_1

- Pour tout $j \in \{1, 2\}$ et $K \in \{3, 4\}$ la construction de S_K^j repose sur la "plug-in rule".
- Étant donné $D_n = (O_{(1)}, \dots, O_{(n)})$ un échantillon d'apprentissage :
 - ▶ on construit $\hat{\eta}_K^j(D_n, .)$ estimateur de η_K^j ;
 - ▶ étant donnée (W, Y_K^j) indépendant de D_n , le classifieur plug-in \hat{S}_K^j est défini par :
$$\hat{S}_K^j(D_n; W, Y_K^j) = \mathbb{1}\{\hat{\eta}_K^j(D_n; W, Y_K^j) \geq \frac{1}{2}\}.$$
- La construction de $\hat{\eta}_K^j(D_n, .)$ est fondée sur une méthode d'agrégation appelée le super-learning (van der Laan et al 2007).

Construction de $\hat{\phi}_1$: notations

- Soit $V \geq 2$ et $(B_v)_{1 \leq v \leq V}$ une partition de $\{1, \dots, n\}$.
- Pour tout $v \in \{1, \dots, V\}$ on note $D_n^v = (O_{(i)})_{i \in B_v}$ et $D_n^{(-v)} = (O_{(i)})_{i \notin B_v}$.
- On définit :

$$P_n^{(v)} = \frac{1}{\text{Card}(B_v)} \sum_{i \in B_v} \delta_{O_{(i)}}$$

$$P_n^{(-v)} = \frac{1}{n - \text{Card}(B_v)} \sum_{i \notin B_v} \delta_{O_{(i)}},$$

- On définit l'estimateur empirique du risque cross-validé \hat{R} et l'oracle \tilde{R} par :

$$\hat{R}(\hat{S}_K^j) = \frac{1}{V} \sum_{v=1}^V R^{(P_n^v)} \left(\hat{S}_K^{j(-v)} \right)$$

$$\tilde{R}(\hat{S}_K^j) = \frac{1}{V} \sum_{v=1}^V R^{(P_0)} \left(\hat{S}_K^{j(-v)} \right),$$

avec

$$R^{(P_n^v)}(\hat{S}_K^{j(-v)}) = \frac{1}{\text{Card}(B_v)} \sum_{i \in B_v} \mathbb{1}\{\hat{S}_K^{j(-v)}(W_i, Y_{K,i}^j) \neq A_i\}.$$

Construction de $\hat{\phi}_1$: principe de la validation croisée V -fold

- L'estimateur $\hat{S}_{\hat{K}}^j$ sélectionné par validation croisée V -fold est défini par :

$$(\hat{K}, \hat{j}) \in \arg \min_{(K, j)} \hat{R}(\hat{S}_K^j).$$

- L'oracle est défini par :

$$(\tilde{K}, \tilde{j}) \in \arg \min_{(K, j)} \tilde{R}(\hat{S}_K^j).$$

- Étant donnée une nouvelle observation $Z = (W, Y_3^1, Y_4^1, Y_3^2, Y_4^2)$, le classifieur cross-validé $\hat{\phi}_1$ est défini par :

$$\hat{\phi}_1(D_n)(Z) = \mathbb{1}\{\hat{S}_{\hat{K}}^j(D_n; Z) \geq \frac{1}{2}\}.$$

- Sous des hypothèses générales, on peut montrer que [Dudoit et van der Laan 2003] :

$$0 \leq E \left[\tilde{R}(\hat{\phi}_1(D_n)) - \mathcal{R}^* \right] \leq E \left[\tilde{R}(\hat{S}_{\hat{K}}^j(D_n)) - \mathcal{R}^* \right] + O \left(\frac{\log(\text{Card}(\mathcal{H}_1))}{\sqrt{np_n}} \right),$$

avec $p_n \simeq \frac{1}{V}$ et $\mathcal{R}^* = \min_{S \in \mathcal{S}} R^{(P_0)}(S)$.

Introduction

Modélisation

Procédure de classification

Application aux données réelles

Perspectives

Annexe

Application de la procédure de classification aux données réelles

- On évalue les performances de la méthode de classification par la règle du leave-one-out :

	$\hat{\phi}_1$	$\hat{\phi}_2$
% bien classé	83 %	80 %

- Résultats de l'extension de la procédure de classification par ajout des anciennes mesures résumées :

	$\hat{\phi}_1$	$\hat{\phi}_2$
% bien classé	83 %	91 %

Introduction

Modélisation

Procédure de classification

Application aux données réelles

Perspectives

Annexe

Conclusions et perspectives

- Conclusion :
 - ▶ modélisation par un processus de diffusion apporte de l'information \Rightarrow étape de modélisation cruciale.
 - ▶ résultats obtenus très satisfaisants 91%
- Perspectives :
 - ▶ prise en compte des informations sur l'orientation des patients durant les protocoles
 \Rightarrow modélisation de $B_{1:N}$ par un processus bi-dimensionnel
 - ▶ propriétés des estimateurs des instants de rupture dans le cadre des diffusions.

Introduction

Modélisation

Procédure de classification

Application aux données réelles

Perspectives

Annexe

Super-learning : introduction

- Le super learning est une procédure de machine learning d'agrégation de prédicteurs fondée sur le principe de la validation croisée.
- L'objectif du super learning est le suivant
 - a) étant donné un échantillon d'apprentissage $\mathcal{L} = \{(X_i, Y_i) : i = 1, \dots, n\}$ où Y_i sont les variables à expliquer et X_i les variables explicatives,
 - b) étant donnée une librairie contenant $K(n)$ algorithmes de prédiction (Modèles linéaires, Forêts aléatoires, Bagging, ...),
 - c) étant donnée une fonction de perte $L : \mathbb{R}^2 \rightarrow \mathbb{R}$ (elle dépend du problème d'intérêt),
 - d) sélectionner la combinaison linéaire convexe de cette librairie minimisant le risque induit par L pour estimer $E[Y|X]$.

Super-learning : principe

Le principe, fondé sur la validation croisée, est le suivant :

1. $\forall k \leq K(n)$ on calcule, à l'aide de \mathcal{L} , $\hat{\psi}_k(W)$ estimateur de $\mathbb{E}[Y|X]$ obtenu avec le $k^{\text{ème}}$ algorithme de prédiction.
2. On divise \mathcal{L} en V blocs $\text{Val}(1), \dots, \text{Val}(V)$. on note $T(\nu) = \mathcal{L} \setminus \text{Val}(\nu)$.
3. $\forall k \leq K(n)$, $\nu \leq V$, on calcule $\hat{\psi}_{k,T(\nu)}(W)$ à partir de $T(\nu)$.
4. $\forall k \leq K(n)$, $\nu \leq V$, $O_i \in \text{Val}(\nu)$, on calcule $\hat{\psi}_{k,T(\nu)}(W_i)$.
5. Finalement on dispose d'une matrice $Z = \{\hat{\psi}_{k,T(\nu)}(W_{\text{Val}(\nu)})\}$ de taille $n \times K(n)$
6. On calcule alors : $\hat{\alpha} = \arg \min_{\alpha} \sum_{i=1}^n L(Y_i, (Z.\alpha)_i)$.
7. On obtient alors l'estimateur par super learning :

$$\hat{\psi}_{SL}(W) = \sum_{k=1}^{K(n)} \hat{\alpha}_k \hat{\psi}_k(W)$$