Report of scientific project METADEB :

Numerical exploration, sensitivity analysis and Bayesian calibration of (meta)models of metabolism from Dynamic Energy Budget (DEB) theory.

Application to the classification of growth stanzas of Indian Ocean Yellowfin tuna (*Thunnus albacares*)

Nicolas Bousquet[*], Emmanuel Chassot[†], Sébastien Da Veiga[‡],
Thierry Klein[*], Bertrand Iooss[*], Agnès Lagnoux[*]

June 9, 2015

1 Presentation

This scientific report adresses the issues which were considered in the METADEB project submitted to the Toulouse IDEX in January, 2014, then supported by the Institut de Mathémetique de Toulouse (IMT) and cofunded by the EMOTION ANR project supported by the Institut de Recherche pour le Développement (IRD) between 2014 and 2015.

This report is written alternatively in English and French.

Contents

[*] Institut de Mathématique de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31400 Toulouse, France.
[†] Institut de Recherche pour le Développement, Seychelles Fishing Authority, Victoria, Mahé Island
[‡] SNECMA Montereau, 77019 Melun, France

2 Problématique

On s'intéresse à une représentation probabiliste de l'évolution du métabolisme de thons tropicaux. Le cas-test utilisé est le *Yellowfin* (albacore) de l'Océan Indien. Selon un vaste ensemble de conditions biologiques et environnementales, fort mal connu, les principes métaboliques peuvent expliquer différents types de croissance en taille et en poids. Différencier ceux-ci en améliorant la connaissance de l'ensemble des conditions constitue la motivation de ce travail. D'une part, la question de l'estimation des paramètres de modèles bioénergétiques et de l'incertitude qui leur est associée est fondamentale car on a tendance actuellement à en fixer beaucoup pour des raisons de commodité. D'autre part, relier des aspects de croissance et de reproduction dans un modèle mécaniste qui explicite comment l'énergie est utilisée et allouée est un enjeu fort des études écologiques, qui visent à comprendre les "tradeoffs" entre survie/croissance/reproduction qui définissent la "fitness" d'une population (capacité à se maintenir) et sa stratégie générale pour faire face à l'environnement (et notamment l'impact de l'exploitation humaine).

2.1 Sources d'information

Notons $L(t)$, $W(t)$, $F(t)$ la longueur, le poids et la fécondité (nombre d'oeuf émis par une femelle) d'un animal à l'âge t. Ces courbes sont croissantes avec t (excepté $F(t)$) et bornées pour des raisons biologiques. Nous disposons de plusieurs sources d'information permettant de construire ces fonctionnelles.

- Des données bibliographiques et *in situ* $\mathbf{D}^* = (L^*, W^*, F^*)$ attribuées à un ensemble d'âges t^* attribué (reconstitué avec un certain bruit connu à partir d'un modèle de lecture d'otolithe) ; ces données proviennent de bibliographies (données larvaires), d'expériences en laboratoire (données de fécondité, (54) et d'expériences de capture-recapture. Voir Dortel et al. (2014) pour un récapitulatif.
- Un simulateur de courbes $\{L(t), W(t), F(t)\}_{t\in I} = g_\theta(X)$ où I est un domaine temporel fixé et :
  1. $X \in \Omega \subset \mathbb{R}^d$ est un vecteur de paramètres aléatoires de loi $f(x)$, correspondant à des paramètres environnementaux et individus-centrés ; la loi $f(x)$ représente donc la distribution des environnements possibles d'un animal en particulier
  2. $\theta \in \Theta \subset \mathbb{R}^q$ est un vecteur de paramètres aléatoires de loi $\pi(\theta)$ dirigeant la croissance type d'un individu de l'espèce considéré ; ils sont par exemple les paramètres d'un modèle de métabolisme issu de la théorie DEB (*Dynamic Energy Budget*).
  3. Quelques informations *a priori* sur $f(x)$ et $\pi(\theta)$ qui proviennent de calages de modèles sur des espèces proches, ou d'expériences en laboratoire.

  Le modèle de simulation pour le *Yellowfin*, qui repose sur des équations aux dérivées partielles modélisant le flux d'énergie dans un individu, est décrit dans Dortel et al. (2014). Voir Figure 1 pour une illustration. En particulier, les paramètres X et θ sont listés dans le tableau 2 de Dortel et al. (2014). Nous en rappelons les principales caractéristiques en Section 3.

Remarque. On différencie θ et X car les paramètres du "code" et ceux des inputs X sont soumis à des incertitudes de nature différente. Il est cependant envisageable qu'une structure de dépendance entre X et θ soit pertinente.

2.2 Objectifs généraux

Les courbes de croissance les plus réalistes présentent des 'stances" (ou points-selles) correspondant à des ralentissements de la croissance (en taille et en poids). Ceux-ci sont supposés correspondre à des instants de vie où l'animal est parvenu à un stade mature et peut devenir reproducteur, puis à un stade

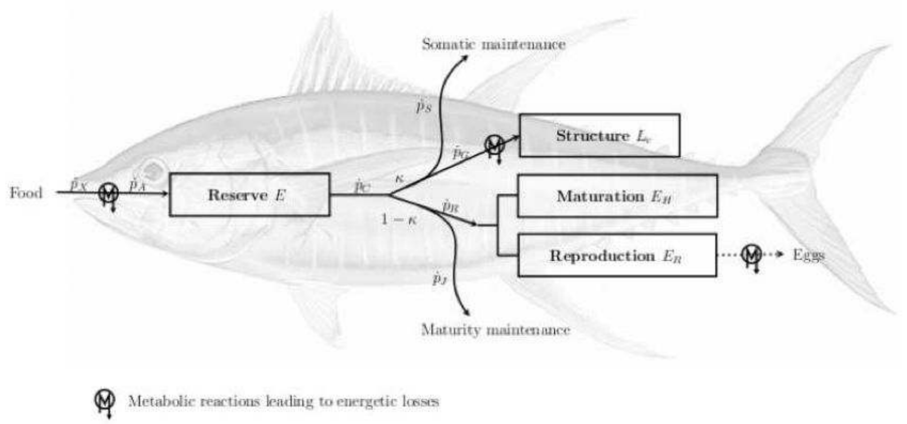


Figure 1: Représentation conceptuelle (tirée de Dortel et al. 2014) d'un modèle DEB pour le *yellowfin*. Les flux d'énergie sont définis dans la Table 1 de Dortel et al. (2014).

de senescence où la vitesse de son métabolisme décroît naturellement. La validation de l'hypothèse d'une ou deux stances (voire trois) et la détermination des conditions biologiques X et gammes de longueurs/poids/âge correspondant constitue des problèmes intéressants pour les scientifiques halieutes. En effet, la compréhension du métabolisme et la relation entre les valeurs de X et les conditions environnementales sont des facteurs déterminants pour la gestion des populations exploitées. En particulier, il s'agit de pouvoir préciser les caractéristiques biologiques des populations ciblées par certains engins de pêche (ex: taille des mailles des filets). En termes de renouvellement de la ressource, l'impact n'est évidemment pas le même lorsque la population ciblée correspond majoritairement à des reproducteurs ou à des juvéniles.

La difficulté principale est la très forte incertitude sur $f(x)$ et $\pi(\theta)$, pour lesquelles on dispose de formes *a priori* assez vagues (par exemple des lois uniformes). Notons $\eta_{x,\theta}$ le nombre de stances associé à une trajectoire bivariée (taille/poids) $g(x)$. Les objectifs sont les suivants.

1. **Classifier le domaine des entrées** $\Omega \times \Theta$ en $(\Omega_1 \times \Theta_1) \cup (\Omega_2 \times \Theta_2) \cup (\Omega_3 \times \Theta_3) \cup \Delta$ tel que

$$\Omega_i \times \Theta_i \quad = \quad \{x \in \Omega,\ \theta \in \Theta,\ \eta_{x,\theta} = i\}$$

   et $\Delta \subset \Omega \times \Theta$ regroupe l'ensemble des autres formes de courbes, considérées comme irréalistes d'un point de vue biologique. On notera qu'il faut définir proprement $\eta_{x,\theta}$ en fonction de des dérivées de $L_x(t)$ et $W_x(t)$ (courbes de longueur et poids associés à la situation biologique x). Ce problème de classification devrait naïvement être mené tous sur les éléments (x, θ) de $\Omega \times \Theta$, ce qui est impossible. Il faut donc trouver une stratégie intelligente d'utilisation du modèle (code) numérique g et du parcours de l'espace $\Omega \times \Theta$. L'utilité de cette démarche est de pouvoir comparer les domaines $\Omega_i \times \Theta_i$ avec les estimations déjà connues pour des espèces proches et d'écarter des domaines irréalistes en regard des connaissances biologiques. Il est important de comprendre quels inputs peuvent générer une courbe qui n'a rien à voir avec 99% des modèles utilisés pour les poissons. Les données observées sont par ailleurs partiales, car certains poissons à croissance rapide, dotés d'une moindre capacité d'association en bancs, leur permettent d'échapper aux pêcheurs (senneurs). Or le modèle DEB ne modélise pas du tout les aspects de comportement social et d'agrégation des thons. Les inputs favorisant des courbes à croissance rapide, doivent être spécifiés et faire l'objet d'un comparatif avec l'expertise IRD, qui récupère d'autres données (filets maillants).

2. **Utiliser les données observées** pour affiner la loi jointe $f(x,\theta)$ sur chacun des domaines $\Omega_i \times \Theta_i$. Formellement, ce problème d'inversion consiste à estimer chaque loi

$$f_i(x,\theta|\mathbf{D}^*) \quad \propto \quad \ell(\mathbf{D}^*|x,\theta)f(x)\mathbb{1}_{\{x\in\Omega_i\}}\pi(\theta)\mathbb{1}_{\{\theta\in\Theta_i\}} \tag{2.1}$$

   où $\ell(\mathbf{D}^*|x,\theta)$ est la vraisemblance des observations *in situ*. Celle-ci est décrite au § 4. L'atteinte de cette loi peut être réalisée au travers d'un algorithme fondé sur les chaînes de Markov, qui nécessite dès lors de très nombreux appels au modèle g. Il semble sans doute intéressant de produire une stratégie tirant parti d'une approximation simple de g.

3. **Mener une analyse de sensibilité** : dans leur domaine de validité et selon la loi f_i trouvée, quel(s) paramètre(s) jouent le plus sur la forme de la courbe et la vitesse de croissance (valeur de la dérivée) ? S'inspirer des travaux préparatoires de Dortel et al. (2014).

4. **Comparer le modèle g calibré** sur chaque $\Omega_i \times \Theta_i$ avec des modèles statistiques de croissance très simplifiés, assez connus dans le monde écologique, comme le modèle VB (Von Bertalanffy) log K (2 stances) de dynamique de longueur :

$$L(t) \quad = \quad L_\infty \left(1 - \exp\left\{-\left[(k_2-k_1)(t-t_1) - k_1(t-t_0)\right]\right\}\right) \tag{2.2}$$

   auquel on peut adjoindre un bruit environnemental (typiquement lognormal). Ces modèles statistiques sont peut-être impuissants à représenter suffisamment finement les variations de la croissance proposée par le simulateur g. Mais ils sont pratiques pour proposer des clés taille-âge et poids-âge faciles à implémenter et rapides d'utilisation dans des modèles de dynamique de population, dont l'objectif ultime est de proposer des points de référence biologiques (tel le rendement maximal durable) et des quotas d'exploitation.

5. L'objectif final est donc de disposer de clés taille-âge et poids-âge $(L(t), W(t))$, tel le modèle simplifié (2.2), associée chacune à une loi $f_i(x,\theta|D^*)$ décrivant des conditions environnementales et biologiques réalistes du point de vue d'une hypothèse de croissance. Ces modèles iront par la suite nourrir des modèles de dynamique des populations plus vastes, possédant certains paramètres de *design* (tels des taux ou efforts de pêche) à tenter d'optimiser.

3 Description of Yellowfin DEB model

Three state variables are defined, the dynamics of which being described in Table 1 :

u the energy amount of the reserve E ;

u the structural length L_v, which is connected through a scale transformation to the fork length L (named F_L in (13), see Equ. (4)) ; 2015

u the maturity level E_H.

The status of the reproduction buffer is quantified by an auxiliary state variable E_R whose dynamics is zero for sexually immature individuals. Cyclically every 6 months the energy stored in the reproduction buffer is converted to eggs. Technically E_R is given the value 0 when a time-dependent indicator $S(t) > 0$, being is defined as a sinusoidal function with a period of 182 days

$$S(t) = \sin\left(2\pi \frac{t}{182}\right).$$

Between each energy conversion, it must be noticed that from the last state equation in Table 1 the dynamics of the auxiliary state variable E_R is linear through time. The energy dimension of the reserve is replaced by a dimensionless quantity:

$$0 \;\leqslant\; e \;=\; \frac{[E]}{[E_m]} \;=\; \frac{E}{L_v^3[E_m]} \;\leqslant\; 1$$

where $[E] = E/L_v^3$ is the reserve density and $[E_m]$ is the maximum reserve density.

The DEB theory considers two environmental forcing variables, the body temperature T (K) and the food density f_X (J.cm^{-3}). The latter is connected to the ingestion rate through a Holling disk equation (Equ. (2) in Dortel et al. 2015). Both variables play important roles the metabolic processes (see Equ. (3) in Dortel et al. 2015). Ingestion and metabolic processes are described respectively in Tables 2 and Tables 3.

Finally, the DEB parameters are described in Table 4. One cannot hope to estimate many of them by capture-recapture data. Therefore some are fixed, following Dortel and al. (2015). For the others, prior information is available under the form of pointwise values obtained for a close fish, the Pacific bluefin tuna (PBT). Note besides that some parameters are lacking to connect the weight and the fork length. But it is considered that $W(t) = \alpha L^\beta(t)$ and that, disposing of numerous couple data, this relationship is known.

State variable	Unit	Dynamic equation
Scaled reserve density	-	$\frac{de}{dt} = (f - e)\frac{\dot{v}}{L_v}$
Structural length	cm	$\frac{dL_v}{dt} = \frac{\dot{v}}{3(e+g)}\left(e - \frac{L_v + L_T}{L_m}\right)$
Maturity level	J	$\frac{dE_H}{dt} = \begin{cases} (1-\kappa)\frac{\{\dot{p}_{Am}\}}{e+g}\left(e\left(g + \frac{L_v + L_T}{L_m}\right)\right)L_v^2 - \dot{k}_J E_H & \text{if} \quad E_H < E_H^p \\ 0 & \text{else} \end{cases}$
Reproduction buffer	J	$\frac{dE_R}{dt} = \begin{cases} 0 & \text{if} \quad E_H < E_H^p \\ (1-\kappa)\frac{\{\dot{p}_{Am}\}}{e+g}\left(e\left(g + \frac{L_v + L_T}{L_m}\right)\right)L_v^2 - \dot{k}_J E_H^p & \text{else} \end{cases}$

Table 1: Dynamics of state variable in the DEB model of yellowfin after metamorphosis (from Dortel et al. 2015).

Definition	Unit	Formulation
Energy investment ratio	-	$g = \frac{\dot{v}[E_G]}{\kappa\{\dot{p}_{Am}\}}$
Structural heating length	cm	$L_T = \frac{\{\dot{p}_T\}}{[\dot{p}_M]}$
Maximum structural length	cm	$L_m = \frac{\kappa\{\dot{p}_{Am}\}}{[\dot{p}_M]}$
Maximum reserve density	cm	$[E_m] = \frac{\{\dot{p}_{Am}\}}{\dot{v}}$
Scaled functional response	-	$f = \frac{f_X}{f_X + f_{X_k}}$ $(0 \leqslant f \leq 1)$

Table 2: List of DEB compound parameters (from Dortel et al. 2015).

Metabolic process	Energy flux
Ingestion	$\dot{p}_X = \{\dot{p}_{Xm}\} f L_v^2 = \frac{\{\dot{p}_{Am}\} f L_v^2}{\kappa_X}$
Assimilation	$\dot{p}_A = \frac{\{\dot{p}_X\}}{\kappa_X} = \{\dot{p}_{Am}\} f L_v^2$
Mobilization	$\dot{p}_C = [E_m] L_v^3 \left(\frac{\dot{v}}{L_v} + \dot{k}_M \left(\frac{L_v + L_T}{L_v} \right) \right) \frac{eg}{e+g}$
Somatic maintenance	$\dot{p}_S = \{\dot{p}_T\} L_v^2 + [\dot{p}_M] L_v^3$
Growth	$\dot{p}_G = \frac{\kappa \dot{p}_C - \dot{p}_S}{[E_G]}$
Maturity maintenance	$\dot{p}_J = \dot{k}_J E_H$
Maturation or reproduction	$\dot{p}_R = (1-\kappa)\dot{p}_C - \dot{p}_J$

Table 3: Energy fluxes (J.d^{-1}) in the yellowfin DEB model (from Dortel et al. 2015).

Definition	Parameter	Unit	PBT	YFT
Primary parameters				
Fraction of mobilized reserve allocated to soma	-	κ	0.7807	0.7807
Surface-area specific maximum assimilation rate	$J.d^{-1}.cm^{-2}$	$\{\dot{p}_{Am}\}$	4783.707	4370.394
Energy conductance	$cm.d^{-1}$	$\dot{v}$	7.056	7.056
Volume-specific somatic maintenance rate	$J.d^{-1}.cm^{-3}$	$[\dot{p}_M]$	17.395	17.395
Surface area-specific somatic maintenance rate	$J.d^{-1}.cm^{-2}$	$\{\dot{p}_T\}$	2215.415	2215.415
Volume-specific costs of structure	$J.cm^{-3}$	$[E_G]$	8563.387	8563.387
Maturity maintenance rate coefficient	d^{-1}	$\dot{k}_J$	0.0612	0.0612
Maturation threshold for metamorphosis	J	E_H^j	6902.209	1368.719
Maturation threshold for end of early juvenile stage	J	E_H^e	969476	192248.7
Maturation threshold for puberty	J	E_H^p	25484395.636	5053598
Fraction of food energy fixed in reserve	-	κ_X	-	0.8
Link parameters				
Shape parameter for juvenile and adult stages	-	δ_j	0.2704	0.2559
Structural volume density	$g.cm^{-3}$	d_v	1	1.0821
Weight-energy coupler	$J.g^{-1}$	ρ_E	7763.975	7346.034
Arrhenius temperature	K	T_A	5298.838	4622.495
Half saturation constant	$J.cm^{-3}$	f_{X_k}	0.0004463	-
Reproduction module parameters				
Energy cost of one egg	J	E_0	3.75	1.295
Fraction of reproduction energy and fixed egg	-	κ_R	0.95	0.95

Table 4: List of DEB parameters and their values for adults Pacific bluefin tuna (PBT) and yellowfin (YFT) at a reference temperature (from Dortel et al. 2015).

In a first approach, by slightly moving the input parameters, a Monte Carlo view of how the DEB model reacts is plotted over Figure 2. Several stanzas can indeed be observed. It must be noticed that the DEB model encompasses a submodel for the food functional response f, which is assumed to be logistic (Figure 3)

During the workshop, prior distributions were elaborated such that the simulation of lengths and weights could be regarded as relevant with respect to the available data.

4 Likelihood of observations

The available observations $\mathbf{D}^*$ are bibliographical results $\mathbf{D}_1^*$ of laboratory experiments and in situ data $\mathbf{D}_2^*$ arising from capture-recapture measurements. Both sources of information are independent. The likelihood of $\mathbf{D}^*$ is then defined by the product of the two likelihoods associated to each source of information:

$$\ell(\mathbf{D}^*|x,\theta) \quad = \quad \ell_1(x,\theta,\sigma)\cdot\ell_2(x,\theta,\tau,\sigma_c,\sigma_r)$$

where the supplementary parameters $\{\sigma,\sigma_c,\sigma_r,\tau\}$ are observational variables.

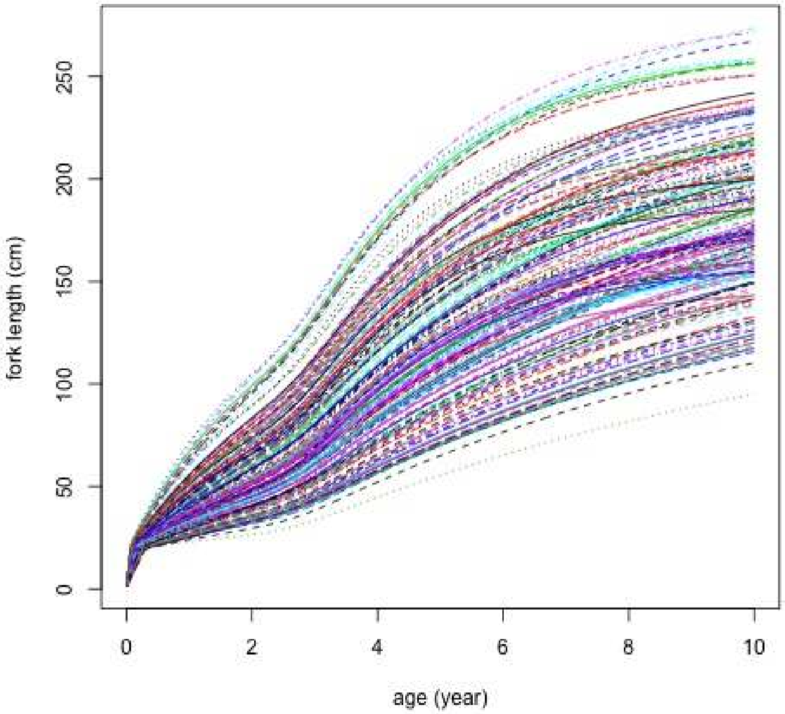


Figure 2: Monte Carlo (basic) simulations of the DEB model for the yellowfin (fork lengths).

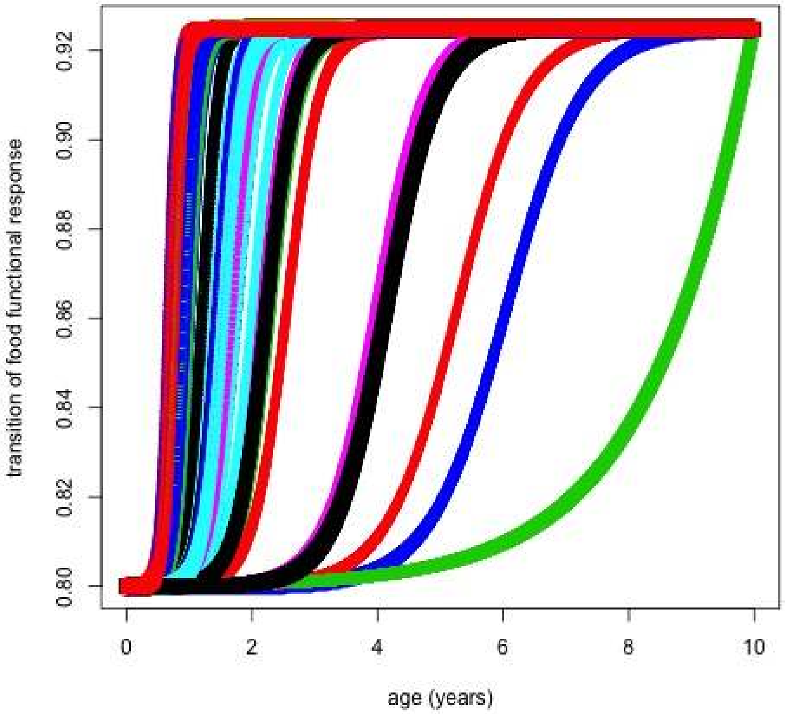
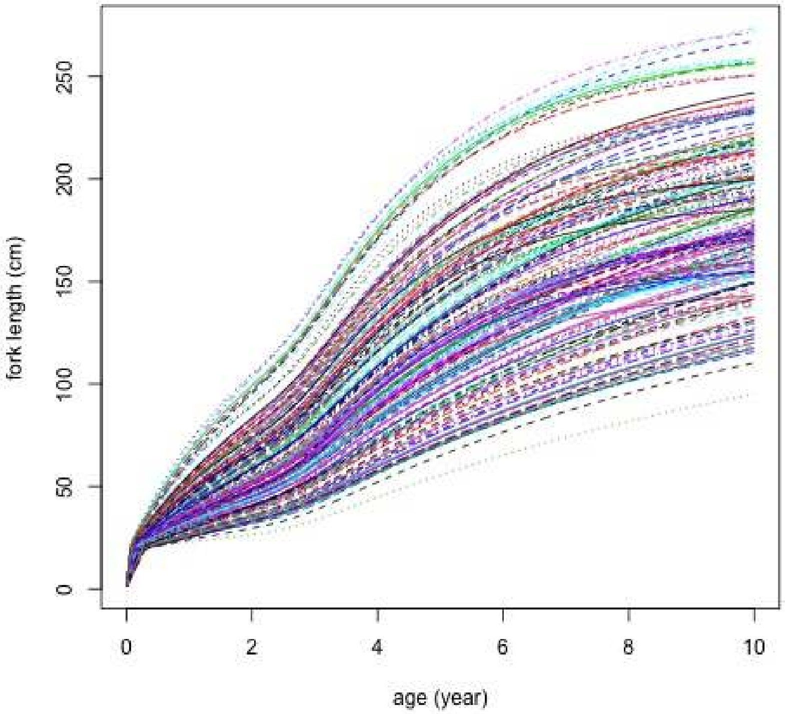


Figure 3: Examples of food response transition curves (from juvenile to adult fish).

Figure 4: Monte Carlo (basic) simulations of the DEB model for the yellowfin (fork lengths).



4.1 Bibliographical larval data

The larval dataset $\mathbf{D}_1^* = \{t_i, L^*(t_i)\}_{i=1,\ldots,n_1}$ where age t_i is know and the length-at-age is observed with error. The following classical hypothesis is done:

$$L^*(t_i) \quad = \quad L(t_i|x,\theta) + \xi_i \;\; \text{with } \xi_i \sim \mathcal{N}(0,\sigma^2), \tag{4.1}$$

where $L(t_i|x,\theta)$ is the output of the numerical model. Their corresponding likelihood is

$$\ell_1(x,\theta,\sigma) \quad = \quad \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left(L^*(t_i) - L(t_i|x,\theta)\right)^2\right\} \tag{4.2}$$

4.2 In situ data

The in situ data $\mathbf{D}^*$ are triplets $\left\{t_j^*, L^*(t_j), L^*(t_j+\Delta_j)\right\}_{j=1,\ldots,n_2}$ corresponding to the obeserved age and to the lengths at capture and recapture, reconstituted using a reading model based on otolithometry (12). A difficulty comes from the fact that the real age t_i of first capture is unknown but estimated by t_i^* because of the reading error. It is assumed that

$$t_j^* \quad = \quad t_j + \nu_j \;\; \text{with } \nu_j \sim \mathcal{N}(0,\tau^2). \tag{4.3}$$

and

$$L^*(t_j) \quad = \quad L(t_j|x,\theta) + \varepsilon_j^c \;\; \text{with } \varepsilon_j^c \sim \mathcal{N}(0,\sigma_c^2), \tag{4.4}$$

$$L^*(t_j+\Delta_j) \quad = \quad L(t_j+\Delta_j|x,\theta) + \varepsilon_j^r \;\; \text{with } \varepsilon_j^r \sim \mathcal{N}(0,\sigma_r^2). \tag{4.5}$$

Fork length measurements are obviously more precise for larger died fish than alive juveniles, hence $\sigma_r \leqslant \sigma_c$. Indeed, measures at first capture are made in stressing conditions (time $<$ 10s) and cannot be repeated while the measures at recapture are conducted on frozen fish with callipers.

The likelihood of $\mathbf{D}_2^*$ can be written under a conditional form:

$$\ell_2(x,\theta,\tau,\sigma_c,\sigma_r) \quad = \quad \prod_{j=1}^{n_2} \left[L^*(t_j+\Delta_j)|L^*(t_j),x,\theta,t_j^*,\tau,\sigma_c,\sigma_r\right]\left[L^*(t_j)|x,\theta,t_j^*,\tau,\sigma_c\right] \tag{4.6}$$

where the symbol [.], popularized by (21), indicates a (cumulative or density) probability distribution. In the rest of the section, each term of (4.6) is detailed to get a fully tractable approximate likelihood which simplifies the computational work needed to reach (2.1) by simulation means.

4.2.1 Approximating the distribution of the length at capture

Consider the second term in (4.6), which corresponds to the law of a first observation (capture). One has

$$\left[L^*(t_j)|x,\theta,t_j^*,\tau,\sigma_c\right] \quad = \quad \int_0^\infty \left[L^*(t_j)|L(t_j|x,\theta),\sigma_c\right]\left[L(t_j|x,\theta)|t_j^*,\tau\right] \; d[L(t_j|x,\theta)].$$

The first integrand term is given by (4.4). An approximation of the second term can be produced by considering that

$$L(t_j|x,\theta) \quad = \quad L(t_j^*|x,\theta) + \sum_{k=1}^{\infty} \frac{(t_j^*-t_j)^k}{k!} L^{(k)}(t_j^*|x,\theta)$$

which exists by smooth regularity of the output at each time (age) step.
A first-order approximation gives (provided $L'(t^*_j|x,\theta) \neq 0$)

$$\left[L(t_j|x,\theta)|t^*_j,\tau\right] \quad \sim \quad \mathcal{N}\left(L(t^*_j|x,\theta), \tau^2\left\{L'(t^*_j|x,\theta)\right\}^2\right)$$

and finally, benefiting from Gaussian conjugation and after some algebraic work, approximately

$$\left[L^*(t_j)|x,\theta,t^*_j,\tau,\sigma_c\right] \quad \propto \quad \frac{\mu_{c,t^*_j}}{\sigma_c\tau\left|L'(t^*_j|x,\theta)\right|}\exp\left(\frac{\eta^2_{c,t^*_j}}{2\mu^2_{c,t^*_j}} - \lambda^2_{c,t^*_j}\right)$$

where the symbol $\propto$ stands for proportionality, and with

$$\begin{aligned}
\mu^2_{c,t} &= \left(\sigma_c^{-2} + \tau^{-2}\left\{L'(t|x,\theta)\right\}^{-2}\right)^{-1}, \\
\eta_{c,t} &= \mu^2_{c,t}\left[\frac{L^*(t)}{\sigma_c^2} + \frac{L(t|x,\theta)}{\tau^2\left\{L'(t|x,\theta)\right\}^2}\right], \\
\lambda^2_{c,t} &= \frac{1}{2}\left(\frac{L^*(t)}{\sigma_c}\right)^2 + \frac{1}{2}\left(\frac{L(t|x,\theta)}{\tau L'(t|x,\theta)}\right)^2.
\end{aligned}$$

4.2.2 Approximating the conditional distribution of the length at recapture

The first term in (4.6) corresponds to the distribution of the second observation (recapture) conditioned on the first observation (capture). Denote

$$\begin{aligned}
\delta^*_t &= L^*(t+\Delta_t) - L^*(t) && \text{the observed difference,} \\
\delta_t(x,\theta) &= L(t+\Delta_t|x,\theta) - L(t|x,\theta) && \text{the simulated difference} \\
\delta'_t(x,\theta) &= L'(t+\Delta_t|x,\theta) - L'(t|x,\theta).
\end{aligned}$$

Since (simplifying the notation $\delta_j = \delta_{t_j}$ and $\delta_{j^*} = \delta_{t^*_j}$)

$$L^*(t_j+\Delta_j) \quad = \quad L^*(t_j) + \delta_j(x,\theta) + \epsilon^c_j + \epsilon^r_j,, \tag{4.7}$$

hence

$$\left[L^*(t_j+\Delta_j)|L^*(t_j),x,\theta,t^*_j,\tau,\sigma^c,\sigma^r\right] \quad = \quad \int_0^\infty \left[L^*(t_j+\Delta_j)|L^*(t_j),\delta_j(x,\theta),\sigma^c,\sigma^r\right] d\left[\delta_j(x,\theta)|t^*_j,\tau\right].$$

The first integrand term is provided by the observational equation (4.7). Based on the same kind of first-order approximation (which is more defensible since at largest sizes length increasing is weak),

$$\delta_j(x,\theta) \quad \simeq \quad \delta_{j^*}(x,\theta) + \left(t^*_j - t_j\right)\delta'_{j^*}(x,\theta),$$

it comes (approximately)

$$\left[\delta_j(x,\theta)|t^*_j,\tau'^2\right] \quad \sim \quad \mathcal{N}\left(\delta_{j^*}(x,\theta), \tau^2\delta'^2_{j^*}(x,\theta)\right).$$

Hence, by conjugation again, it can be found

$$\begin{aligned}
\left[L^*(t_j+\Delta_j)|L^*(t_j),x,\theta,t^*_j,\tau,\sigma_c,\sigma_r\right] \quad &\propto \quad \int_0^\infty \frac{1}{\sqrt{\sigma_c^2+\sigma_r^2}}\exp\left(-\frac{1}{2(\sigma_c^2+\sigma_r^2)}\left\{\delta^*_j - \delta_j(x,\theta)\right\}^2\right) \\
&\qquad \times \frac{1}{\tau\left|\delta'_{j^*}(x,\theta)\right|}\exp\left(-\frac{1}{2\tau^2\delta'^2_{j^*}(x,\theta)}\left\{\delta_j(x,\theta) - \delta_{j^*}(x,\theta)\right\}^2\right), \\
&\propto \quad \frac{\mu_{r,t^*_j}}{\sqrt{\sigma_c^2+\sigma_r^2}\tau\left|\delta'_{j^*}(x,\theta)\right|}\exp\left(\frac{\nu^2_{r,t^*_j}}{2\mu^2_{r,t^*_j}} - \lambda^2_{r,t^*_j}\right)
\end{aligned}$$

with

$$
\begin{aligned}
\mu^2_{r,t} &= \left(\left[\sigma_c^2 + \sigma_r^2\right]^{-1} + \tau^{-2} \left\{\delta'_t(x,\theta)\right\}^{-2} \right)^{-1}, \\
\eta_{r,t} &= \mu^2_{c,t} \left[\frac{\delta^*_t}{\sigma_c^2 + \sigma_r^2} + \frac{\delta_t(x,\theta)}{\tau^2 \delta'^2_t(x,\theta)} \right], \\
\lambda^2_{r,t} &= \frac{1}{2} \frac{(\delta^*_t)^2}{\sigma_c^2 + \sigma_r^2} + \frac{1}{2} \left(\frac{\delta_t(x,\theta)}{\tau \delta'_t(x,\theta)} \right)^2 .
\end{aligned}
$$

5 Bayesian calibration of input distributions

The Bayesian actualisation principle is carried out in the following Gibbs sampler, each step of which being conducted using an usual Hastings-Metropolis algorithm. A classical choice of random walk is made for the instrumental distribution $\rho(m, \sigma^2)$ with mean m and variance σ^2, with decreasing innovation $\lambda > 0$ (towards a strictly positive limit), such that the algorithm produces an adaptive Markov chain converging almost surely towards the target posterior joint distribution (43).

```
Step 0:
    Sample $X^{(0)} \sim f(X)$ and $\theta^{(0)} \sim \pi(\theta)$
Step i+1: (pour $i \geq 0$)
    1. For $k = 1, \ldots, d$
        (a) sample $\tilde{x}_k^{(i)} \sim \rho_x(x_{k-1}^{(i)}, \lambda_k\{x_{k-1}^{(i)}\}^2)$;
        (b) compute the ratio
            $$\alpha_{i,k} = \frac{\ell\left(\mathbf{D}^*|x_1^{(i-1)},\ldots,x_{k-1}^{(i-1)},\tilde{x}_k^{(i)},x_{k+1}^{(i-1)},\ldots,x_d^{(i-1)},\theta^{(i-1)}\right)}{\ell\left(\mathbf{D}^*|x_1^{(i-1)},\ldots,x_{k-1}^{(i-1)},x_k^{(i-1)},x_{k+1}^{(i-1)},\ldots,x_d^{(i-1)},\theta^{(i-1)}\right)} \frac{f_k\left(\tilde{x}_k^{(i)}\right)\rho_x\left(x_{k-1}^{(i)},\lambda_k\{x_{k-1}^{(i)}\}^2\right)}{f_k\left(x_{k-1}^{(i)}\right)\rho_x\left(\tilde{x}_{k-1}^{(i)},\lambda_k\{\tilde{x}_k^{(i)}\}^2\right)}$$
        (c) select $x_k^{(i)} = \tilde{x}_k^{(i)}$ with probability $\min(1, \alpha_{i,k})$ else fix $x_k^{(i)} = x_{k-1}^{(i)}$.
    2. For $k = 1, \ldots, q$
    3. sample $\tilde{\theta}_k^{(i)} \sim \rho_\theta(\theta_{k-1}^{(i)}, \lambda'_k\{\theta_{k-1}^{(i)}\}^2)$;
    4. compute the ratio
        $$\beta_{i,k} = \frac{\ell\left(\mathbf{D}^*|x^{(i)},\theta_1^{(i-1)},\ldots,\theta_{k-1}^{(i-1)},\tilde{\theta}_k^{(i)},\theta_{k+1}^{(i-1)},\ldots,\theta_d^{(i-1)}\right)}{\ell\left(\mathbf{D}^*|x^{(i)},\theta_1^{(i-1)},\ldots,\theta_{k-1}^{(i-1)},\theta_k^{(i-1)},\theta_{k+1}^{(i-1)},\ldots,\theta_d^{(i-1)}\right)} \frac{\pi_k\left(\tilde{\theta}_k^{(i)}\right)\rho_\theta\left(\theta_{k-1}^{(i)},\lambda'_k\{\theta_{k-1}^{(i)}\}^2\right)}{\pi_k\left(\theta_{k-1}^{(i)}\right)\rho_\theta\left(\tilde{\theta}_{k-1}^{(i)},\lambda'_k\{\tilde{\theta}_k^{(i)}\}^2\right)}$$
    5. select $\theta_k^{(i)} = \tilde{\theta}_k^{(i)}$ with probability $\min(1, \beta_{i,k})$ else fix $\theta_k^{(i)} = \theta_{k-1}^{(i)}$.
```

In practice, a maximum likelihood estimation of (x, θ) is conducted to initiate the calculation, then three independent Markov chains are sampled in parallel, such that the usual convergence diagnostics (e.g., Brooks-Gelman multivariate statistics (9) ; a recent survey is made in (45)) be used to select the burn-in period and detect the stationarity of chains sampling within the target posterior joint distribution. A

Gibbs block algorithm can be used for sampling the x all together. A generic implementation of this algorithm was made, and a typical behavior is plotted on Figure 5 (based on data simulated from the toy model).

6 Classification

7 Posterior sensitivity study

A sensitivity study is conducted on the DEB model restricted to the inputs described in Section 3. This study is adapted to the functional nature of the outputs of the DEB numerical model by means of a generalisation of well-known Sobol indices. The technical principle of the generalization is described in Appendix A.

To conduct such a sensitivity analysis, prior distributions on input parameters are (a minima) required. The choice of these distribution is detailed in next paragraph.

7.1 Eliciting prior distributions

The available prior knowledge consists in single values of input parameters, arising from past calibrations conducted on a close species: the Pacific bluefin tuna (PBT).

8 Numerical experiments

8.1 Toy model

The global sensitivity analysis was firstly tested on a toy example, which is based on the so-called Von Bertalanffy (VB) log K growth curve (36, 26). This model allows a smooth transition between two different growth rate coefficients (k_1 and k_2) through modeling changes in growth by a logistic function. According to this model, the expected fork length at age t is expressed as:

$$L(t) \quad = \quad L_\infty \left\{ 1 - \exp\left(-k_2(t-a_0)\right) \left[\frac{1+\exp\left(-\beta(t-a_0-\alpha)\right)}{1+\exp(\beta\alpha)} \right] \right\}.$$

where L_∞ is the asymptotic length, i.e. the maximum length that fish can reach, α is the inflection point, i.e the relative age to a_0 at which the change in growth occurs and β is the parameter that controls the rate of transition between k_1 and k_2. If $k_2 \ll k_1$, the growth curve presents two clear stanza. It reduces to the usual VB (one-stanza) curve if $k_1 = k_2$. Following (14), a three-stanza curve can be build by extending the VB log K model in

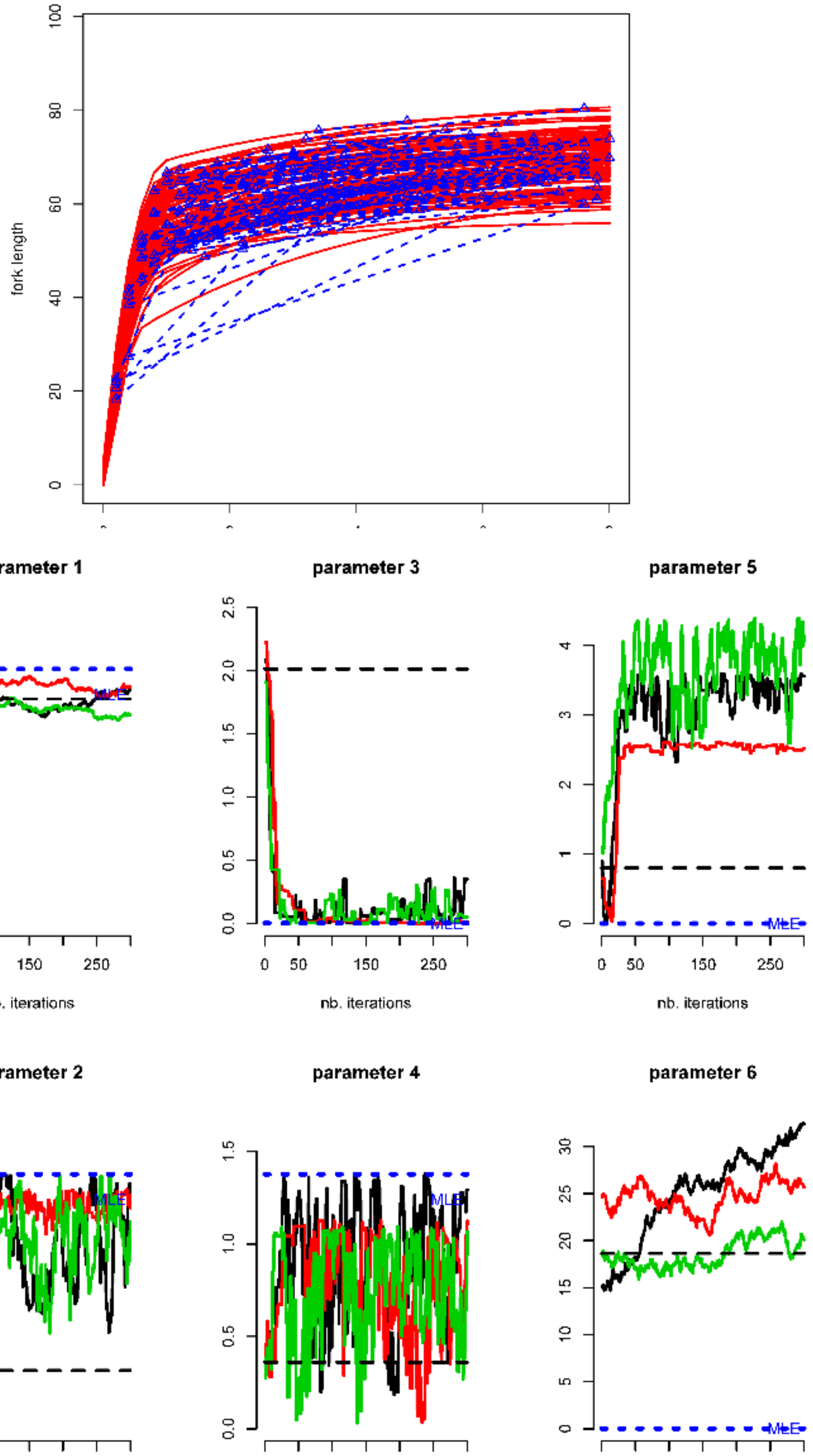
$$L(t) \quad = \quad \frac{L_d}{1+\exp(-k_0(t-a_0))} + (L_\infty - L_d) \left\{ 1 - \exp\left(-k_2(t-a_0)\right) \left[\frac{1+\exp\left(-\beta(t-a_0-\alpha)\right)}{1+\exp(\beta\alpha)} \right] \right\}.$$

where L_d is a transition length, appearing at age $t_d > t_0$, before which the growth is guided by a rate $k_0 < k_2$.

8.1.1 Modelling

A set of input distributions (truncated Gaussian) was considered based on the results obtained for the skipjack tuna by (27). They are summarized in Table 5. An example of 30 curves (10 per each number of stanza) is plotted on Figure 6.

Figure 5: Typical simulated data and starting plot of MCMC parallel chains (two-stanza toy model)



Parameter	L_∞ (cm)	a_0 (y)	α	β	k_1	k_2	k_0	L_d (cm)	t_d (y)
Mean value	70.5	0	0.8	18.9	2	0.35	0.2	20	4
Standard deviation	4.9	0.1	0.15	4.14	0.2	0.08	0.05	4.5	0.3

Table 5: Input distributions for the toy model (in the encompassing three-stanza case), coming from typical values found for the Indian Ocean skipjack tuna.

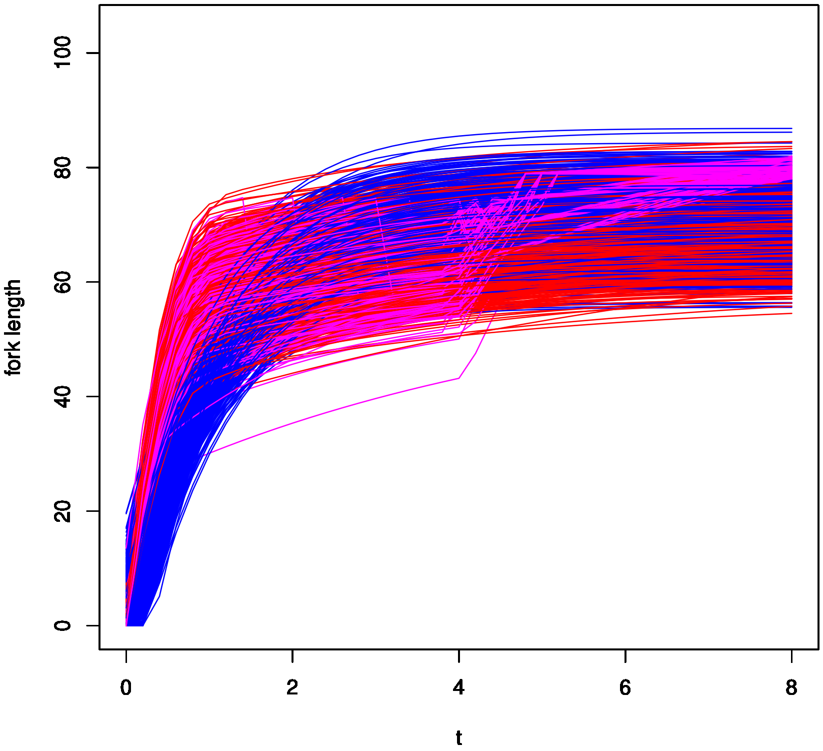


Figure 6: Multi-stanza growth curves plotted using perturbated Von Bertalanffy growth curves.

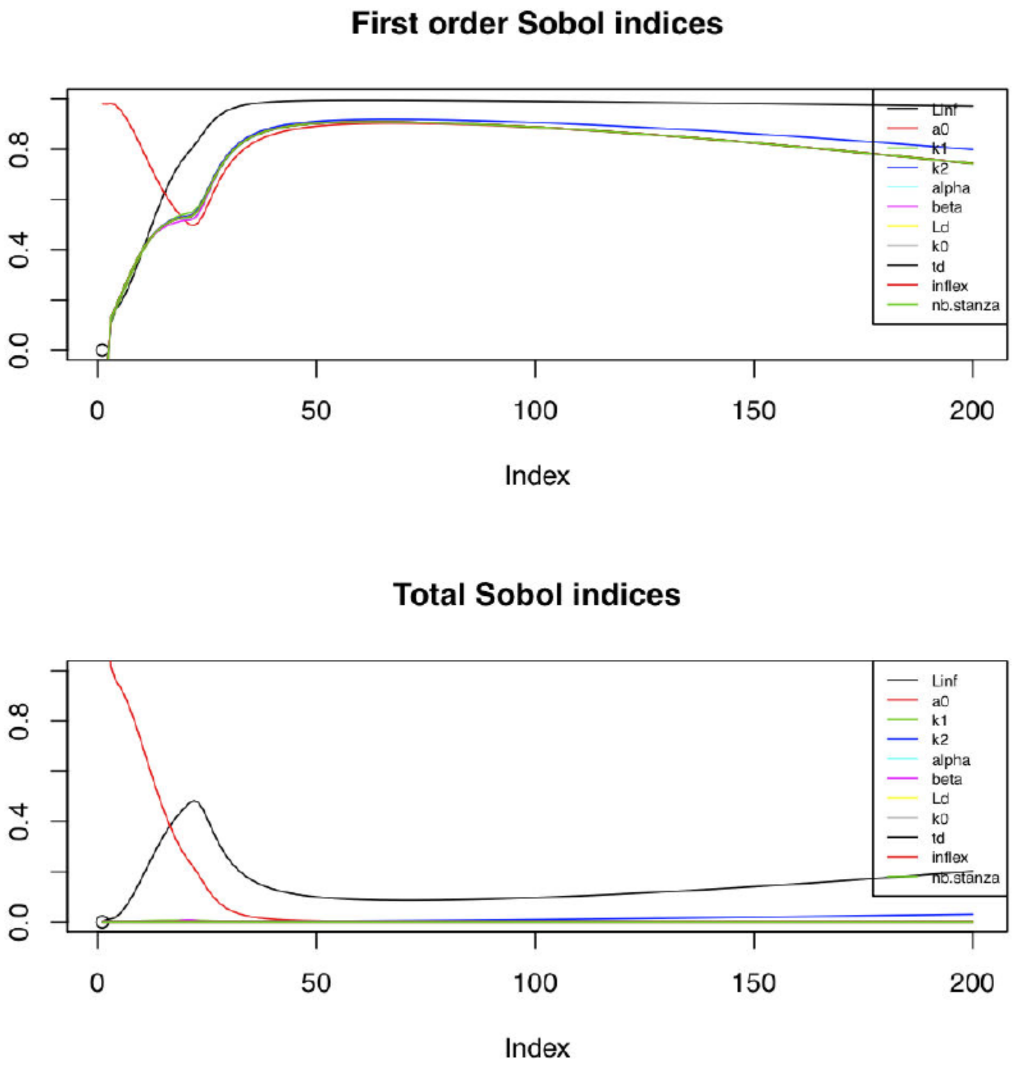


Figure 7: First-order and total functional Sobol indices for the toy model. The parameter "inflex" is a supplementary parameter without impact on the model. The index represents the age (in quarters).

8.1.2 Sensitivity analysis

A Sobol sensitivity analysis is conducted over the toy model. The results obtained for the first-order, total and cumulative Sobol indices are plotted on Figure ?? and 8. See the Appendix for more explanations. Among others, these results highlight the increasing importance of L_{inf} with the age, which is obvious since L_{inf} is the size of oldest fish. An interesting feature is the variation of the importance of a_0.

8.1.3 Classification

Classifying the curves in function of the number of stanza would theoretically require specific developments of distances or divergences between the curves. However, it appears natural to test common things from the literature to establish the maximal number of stanza, then used usual classification tools as $k-$means to go more into depth in the explanations of the appearance of such stanza. A first, logical tool is principal component analysis (PCA). A first illustration of the results of an usual vectorial PCA (not functional PCA), based on a large number of simulated curves, is displayed on Figure 9. These results were obtained using the R function *prcomp*. Three clusters of curves indeed appear by reading the

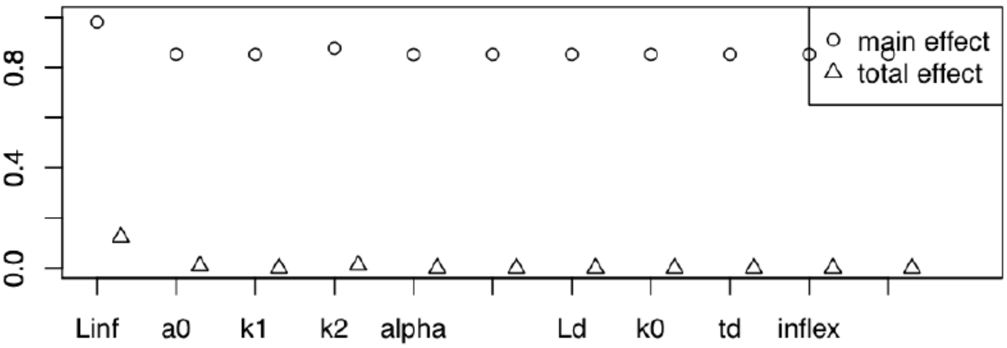


Figure 8: Cumulative Sobol indices for the toy model. The parameter "inflex" is a supplementary parameter without impact on the model. The index represents the age (in quarters).

correlation results between the first dimensions. The last graph on Figure 9 shows that only 4 components are explaining the most part of the variance.

8.2 Motivating case-study: Indian Ocean yellowfin

8.2.1 Capture-recapture data

7432 yellowfin capture-recapture data were collected throughout the Regional Tuna Tagging Project (RTTP-IO), a large scale mark-recapture program supervised by the Indian Ocean Tuna Commission (IOTC), under the authority of the Food and Agriculture Organization (FAO), between 2001 and 2011. The increasing relationship between the time at liberty (Δ_t) and the observed lengths is plotted on Figure 10, while the couples of captured and recaptured lengths are plotted on Figure 11. As it can be observed, for several observations (5.58%) the length at capture is found to be bigger than the length at recapture, which is not feasible biologically. Therefore these observations must be discarded. The discarding rule is not only based on the negativity or nullity of length increasings ; fish presenting an average increasing rate of more than 8 cm per month are assumed not to be realistic (Figure 12). Finally, 7121 couples of capture-recapture observations are selected (plotted on Figure 13).

References

[1] M. Baucells and E. Borgonovo. Invariant probabilistic sensitivity analysis. *Mangmnt Sci.*, pages 2536 – 2549, 2013.

[2] E Borgonovo. A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6):771 784, 2007.

[3] E Borgonovo. A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6):771 784, 2007.

[4] E. Borgonovo, G.B.. Hazen, and E. Plischke. Probabilistic sensitivity measures: Foundations and estimation. *Operations Research*, 2013.

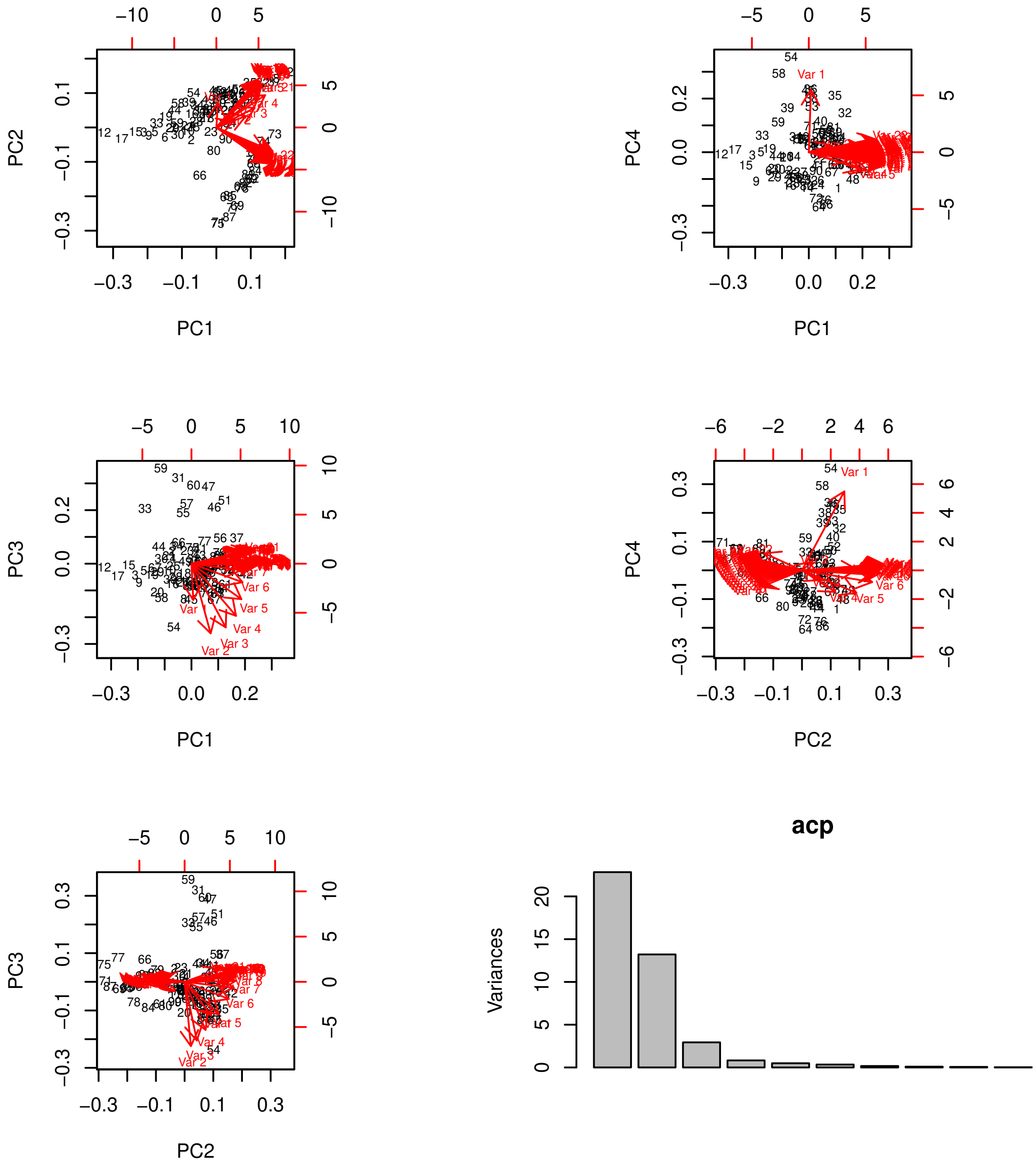


Figure 9: Results of a vectorial principal component analysis (PCA) trying to highlight the divergences between 15,000 growth curves (5,000 are simulated for each sampling model.)

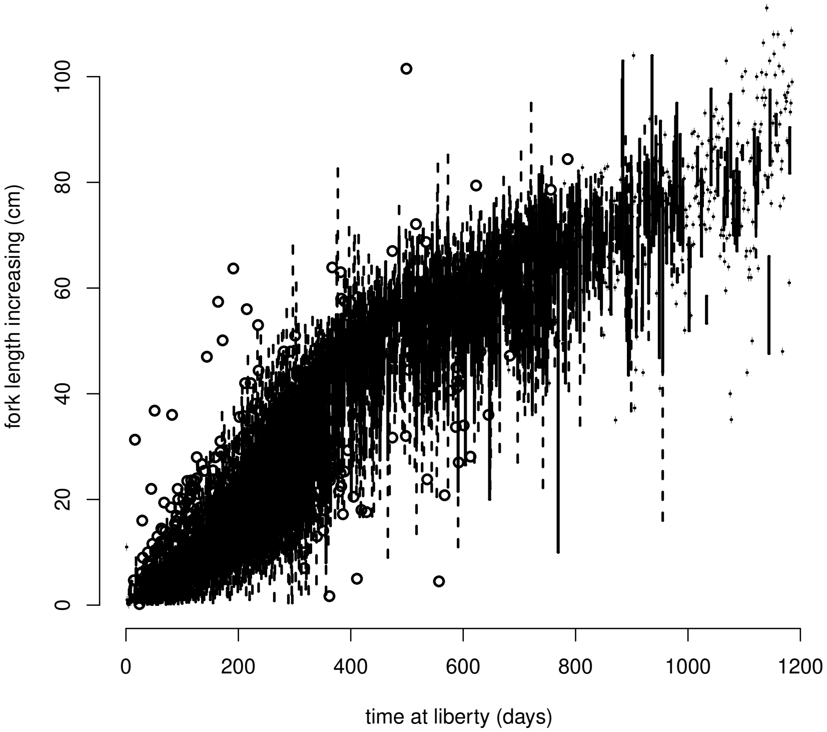


Figure 10: Empirical growth in function of the time at liberty Δ_t between capture and recapture. The term "quarters" refer to quarters of year (trimesters).

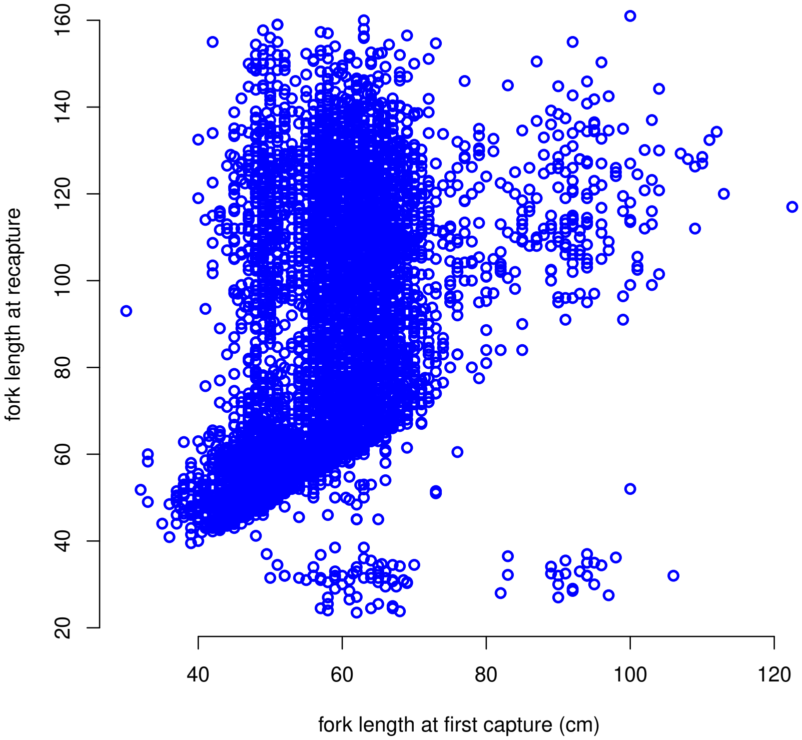


Figure 11: 7432 couples of captured and recaptured fork lengths. In the lower part of the graph, strange observations should be discarded from the assessment.

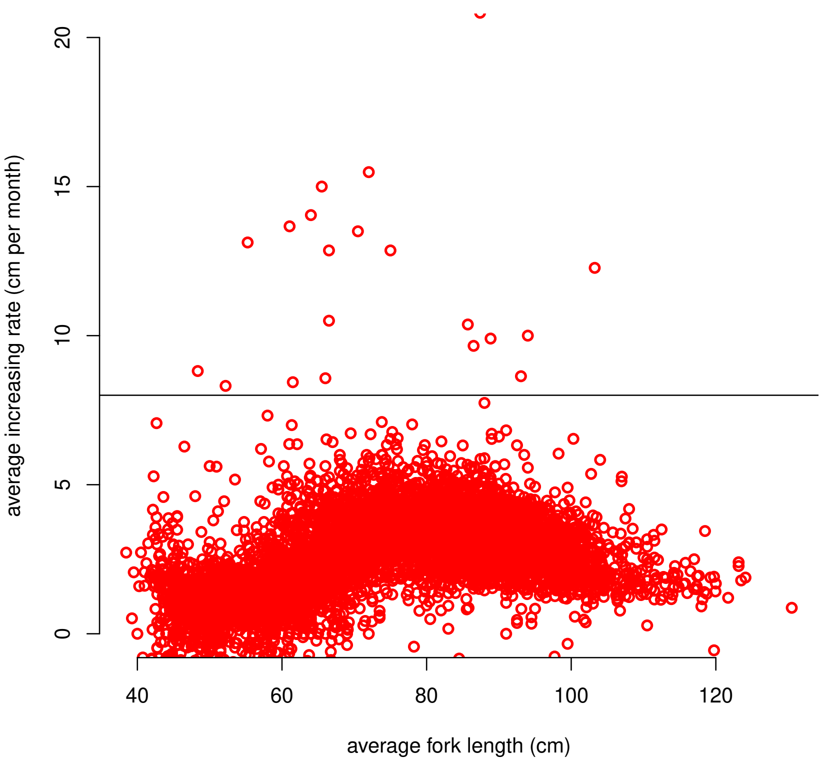


Figure 12: Average increasing rate per month (in cm) and relevance limit.

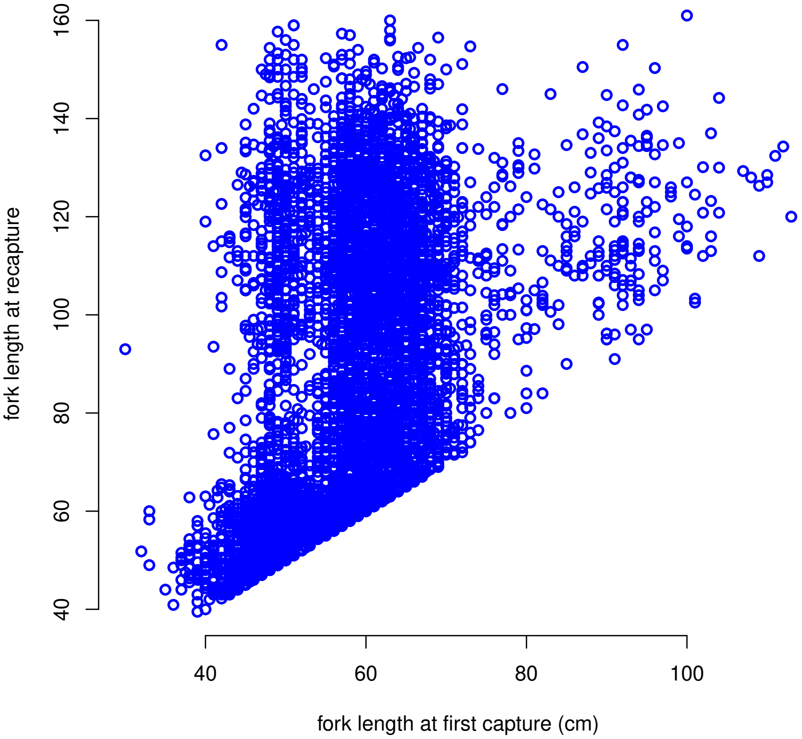


Figure 13: Finally selected couples (7121) of capture-recapture fork lengths.

[5] E. Borgonovo and B. Iooss. Moment independent importance measures and a common rationale. *Submitted*, 2014.

[6] Emanuele Borgonovo, William Castaings, and Stefano Tarantola. Moment independent importance measures: New results and analytical test cases. *Risk Analysis*, 31(3):404–428, 2011.

[7] Emmanuelle Borgonovo, William Castaings, and Stefano Tarantola. Model emulation and moment-independent sensitivity analysis: An application to environmental modelling. *Environmental Modelling & Software*, 34:105–115, 2012.

[8] Charles Bouveyron and Julien Jacques. Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Anal. Classif.*, 5(4):281–300, 2011.

[9] S.P. Brooks and A. Gelman. Fecundity regulation strategy of the yellowfin tuna (thunnus albacares) in the western indian ocean. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[10] Sebastien Da Veiga. Global sensitivity analysis with dependence measures. *J. Stat. Comput. Simul.*, 85(7):1283–1305, 2015.

[11] E. De Rocquigny, N. Devictor, and S. Tarantola. *Uncertainty in industrial practice.* Wiley Online Library, 2008.

[12] E. Dortel, F. Massot-Granier, E. Rivot, J. Million, J.-P. Hallier, E. Morize, J.-M. Munaron, N. Bousquet, and E. Chassot. Accounting for age uncertainty in growth modeling, the case study of yellowfin tuna (*thunnus albacares*) of the indian ocean. *PLOS One*, 2013.

[13] E. Dortel, L. Pecquerie, and E. Chassot. A dynamic energy budget modelling approach to investigate the eco-physiological factors behind the two-stanza growth of yellowfin tuna (submitted). *Fisheries Research*, 2015.

[14] E. Dortel, F. Sardenne, G. Le Crozier, N. Bousquet, and E. Chassot. A three-stanza growth model for indian ocean yellowfin tuna. *(work-in-progress)*, 2015.

[15] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

[16] Frédéric Ferraty and Yves Romain. *The Oxford Handbook of Functional Data Analysis.* Oxford Handbooks in Mathematics, 2010.

[17] Louis Ferré and Nathalie Villa. Multilayer perceptron with functional inputs: an inverse regression approach. *Scand. J. Statist.*, 33(4):807–823, 2006.

[18] J.-C. Fort, T. Klein, and N. Rachdi. New sensitivity analysis subordinated to a contrast. *ArXiv e-prints - Accepted at Communications in Statistics - Theory and Methods*, May 2014.

[19] S. Gaffney. Probabilistic curve-aligned clustering and prediction with mixture models. *PhD thesis. Department of Computer Science. University of California, Irvine, USA*, 2001.

[20] Fabrice Gamboa, Alexandre Janon, Thierry Klein, and Agnès Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electron. J. Stat.*, 8(1):575–603, 2014.

[21] A.E. Gelfand and A. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[22] M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard. Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69(1):31–40, 2013.

[23] Nicholas A. Heard, Christopher C. Holmes, and David A. Stephens. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc.*, 101(473):18–29, 2006.

[24] G. Hébrail, B. Hugueney, Y. Lechevallier, and F. Rossi. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing.*, 73(7-9):1125–1141, 2010.

[25] Francesca Ieva, Anna M. Paganoni, Davide Pigoli, and Valeria Vitelli. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 62(3):401–418, 2013.

[26] P. J. Eveson, G. Laslett, and T. Polacheck. An integrated model for growth incorporating tag-recapture, length-frequency, and direct aging data. *Canadian Journal of Fisheries and Aquatic Sciences*, 61:292–306, 2004.

[27] P. J. Eveson, J. Million, F. Sardenne, and G. Le Croizier. Estimating growth of tropical tunas in the indian ocean using tag-recapture data and otolith-based age estimates. *Fisheries Research*, 163:58–68, 2015.

[28] Julien Jacques and Cristian Preda. Funclust: a curves clustering method using functional random variable density approximation. *Neurocomputing. In press.*, 2013.

[29] Julien Jacques and Cristian Preda. Functional data clustering: a survey. *Adv. Data Anal. Classif.*, 8(3):231–255, 2014.

[30] Gareth M. James and Catherine A. Sugar. Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.*, 98(462):397–408, 2003.

[31] Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364, 1 2014.

[32] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.*, 10:1391–1445, 2009.

[33] Mitsunori Kayano, Koji Dozono, and Sadanori Konishi. Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. *J. Classification*, 27(2):211–230, 2010.

[34] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E (3)*, 69(6):066138, 16, 2004.

[35] B. Krzykacz-Hausmann. Epistemic sensitivity analysis based on the concept of entropy. *SAM0'2001*, 2001.

[36] G. Laslett, P. J. Eveson, and T. Polacheck. A flexible maximum likelihood approach for fitting growth curves to tag-recapture data. *Canadian Journal of Fisheries and Aquatic Sciences*, 59:976–986, 2002.

[37] H. Monod, C. Naud, and D. Makowski. Uncertainty and sensitivity analysis for crop models. In D. Wallach, D. Makowski, and J. W. Jones, editors, *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications*, chapter 4, pages 55–99. Elsevier, 2006.

[38] A. Owen, J. Dick, and S. Chen. Higher order Sobol' indices. *ArXiv e-prints*, June 2013.

[39] A. B. Owen. Variance components and generalized Sobol' indices. *ArXiv e-prints*, May 2012.

[40] Art B Owen. Better estimation of small sobol'sensitivity indices. *arXiv preprint arXiv:1204.4763*, 2012.

[41] Art B Owen. Variance components and generalized sobol' indices. Preprint available at `http://arxiv.org/abs/1205.1774`, 2012.

[42] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[43] G. Robert, C.P.and Casella. *Monte Carlo Statistical Methods (2nd edition)*. New Work: Wiley., 2004.

[44] Fabrice Rossi and Nathalie Villa. Recent advances in the use of SVM for functional data classification. pages 273–280, 2008.

[45] K. Sahlin. *Estimating convergence of Markov chain Monte Carlo simulations*. Ms. in Mathematical Statistics, Stockholm University, 2011.

[46] A. Saltelli, K. Chan, and E.M. Scott. *Sensitivity analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000.

[47] Allou Samé, Faicel Chamroukhi, Gérard Govaert, and Patrice Aknin. Model-based clustering and segmentation of time series with changes in regime. *Adv. Data Anal. Classif.*, 5(4):301–321, 2011.

[48] T. J. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, 2003.

[49] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment*, 1(4):407–414 (1995), 1993.

[50] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, Cambridge, 2012. With a foreword by Thomas G. Dietterich.

[51] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. *J. Mach. Learn. Res.*, 4:5–20, 2008.

[52] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[53] Michio Yamamoto. Clustering of functional data in a low-dimensional subspace. *Adv. Data Anal. Classif.*, 6(3):219–247, 2012.

[54] I. Zudaire, H. Murua, M. Grande, M. Korta, H. Arrizabalaga, JJ. Areso, and A. Delgado-Molina. Fecundity regulation strategy of the yellowfin tuna (thunnus albacares) in the western indian ocean. *Fisheries Research*, 138:80–88, 2013.

A A survey on sensitivity analysis based on order-two measures

See (20) for more details.

A very classical problem in the study of computer code experiments (see (48)) is the evaluation of the relative influence of the input variables on some numerical result obtained by a computer code. This study is usually called sensitivity analysis in this paradigm and has been widely assessed (see for example (49), (46), (11) and references therein). More precisely, the result of the numerical code Y is seen as a function of the vector of the distributed input $(X_r)_{r=1,\cdots,p}$ ($p \in \mathbb{N}^*$). Statistically speaking, we are dealing here with the unnoisy non parametric model

$$Y = f(X_1, \dots, X_p), \tag{A.1}$$

where f is a regular unknown numerical function on the state space $E_1 \times E_2 \times \ldots \times E_d$ on which the distributed variables $(X_1, \dots, X_d)$ are living. Generally, the random inputs are assumed to be **independent** and a sensitivity analysis is performed by using the so-called Hoeffding decomposition (see (52) and (**?**)). In this functional decomposition, f is expanded as an L^2-sum of uncorrelated functions involving only a part of the random inputs. For any subset u of $I_p = \{1, \dots, p\}$, this leads to an index called the Sobol index ((49)) that measures the amount of *randomness* of Y carried in the subset of input variables $(X_i)_{i\in u}$.
More precisely, for any subset u of $I_p := \{1, \dots, p\}$, we denote by $\sim u$ its complement in $\{1, \dots, p\}$. Further, we set $X_u = (X_i, i \in u)$ and $E_u = \prod_{i\in u} E_i$. Then we write

$$Y = f(X) = c + f_u(X_{\mathbf{u}}) + f_{\sim u}(X_{\sim\mathbf{u}}) + f_{u,\sim u}(X_{\mathbf{u}}, X_{\sim\mathbf{u}}), \tag{A.2}$$

where $c \in \mathbb{H}$, $f_u : E_u \to \mathbb{H}$, $f_{\sim u} : E_{\sim u} \to \mathbb{H}$ and $f_{u,\sim u} : E \to \mathbb{H}$ are given by

$$c = \mathbb{E}(Y),\ f_u = \mathbb{E}(Y|X_{\mathbf{u}}) - c,\ f_{\sim u} = \mathbb{E}(Y|X_{\sim\mathbf{u}}) - c,\ f_{u,\sim u} = Y - f_u - f_{\sim u} - c.$$

Thanks to L^2-orthogonality, it remains to take the variance if Y is scalar or the covariance matrix if Y is multivariate of both sides of (A.2) to define the Sobol indices.

In the next subsections, we investigate different situations (Y scalar, multivariate or functional), define for each one the Sobol indices, propose estimators and study their asymptotic properties. Since nothing has been assumed on the nature of the inputs, one can consider the vector $(X_i)_{i\in v}$ as a single input. Thus without loss of generality, let us consider in the sequel the case where u reduces to a singleton.

A.1 Sobol indices for scalar outputs

A.1.1 Black box model and Sobol indices

Here we consider a real output Y square integrable and non deterministic ($\mathrm{Var}(Y) \neq 0$). Taking the variance of both sides of equation (A.2), we are lead to

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y|X_u)) + \mathrm{Var}(\mathbb{E}(Y|X_{\sim u})) + \mathrm{Var}(f_{u,\sim u}(X_{\mathbf{u}}, X_{\sim\mathbf{u}})). \tag{A.3}$$

and (see (46)) a natural way to define the closed Sobol index with respect to u is then

$$S^u_{\mathrm{Cl}} := \frac{\mathrm{Var}(\mathbb{E}(Y|X_u))}{\mathrm{Var}(Y)}.$$

Properties

1. The Sobol indices sum up to 1
$$S^u_{\mathrm{Cl}} + S^{\sim u}_{\mathrm{Cl}} + S^{u,\sim u}_{\mathrm{Cl}} = 1. \tag{A.4}$$
2. $0 \leqslant S^u_{\mathrm{Cl}} \leqslant 1$.
3. S^u_{Cl} is invariant by left-composition of f by any isometry.
4. S^u_{Cl} is invariant by left-composition by any nonzero scaling of f i.e.
$$\text{for any } \lambda \in \mathbb{R}, \;\; S^u_{\mathrm{Cl}}(\lambda f) = S^u_{\mathrm{Cl}}(f);$$

Variational formulation
Using a classical regression result, we see that

$$S^u_{\mathrm{Cl}} = \operatorname*{argmin}_{a \in \mathbb{R}} \left\{ \mathbb{E}\left((Y^u - \mathbb{E}(Y^u)) - a(Y - \mathbb{E}(Y)) \right)^2 \right\}. \tag{A.5}$$

A.1.2 Monte Carlo estimation of $S^u(f)$: Sobol pick freeze method

For X and for any index u of I_p, we define X^u such that $X^u_u = X_u$ and $X^u_i = X'_i$ if $i \neq u$ where X'_i is an independent copy of X_i. We then set

$$Y^u := f(X^u).$$

Notation: From now on, we will denote $\mathrm{Var}(Y)$ by V, $\mathrm{Cov}(Y, Y^u)$ by C_u and $\overline{Z}_N$ the empirical mean of any N-sample $(Z_1, \ldots, Z_N)$ of Z.

The Sobol pick freeze method to estimate S^u_{Cl} relies on the following identity

$$\mathrm{Var}(\mathbb{E}(Y|X_u)) = \mathrm{Cov}\left(Y, Y^u\right). \tag{A.6}$$

that shows how to express S^u_{Cl} in terms of a covariance and leads to a natural estimator.

Design of experiments and estimation phase

We take N independent copies $Y_1, \ldots, Y_N$ (resp. $Y^u_1, \ldots, Y^u_N$) of Y (resp. Y^u). We then use all the available information and define for any $i = 1, \ldots, N$,

$$Z^u_i = \frac{1}{k+1}\left(Y_i + \sum_{j=1}^{k} Y_i^{u_j} \right), \quad M^u_i = \frac{1}{k+1}\left(Y_i^2 + \sum_{j=1}^{k} (Y_i^{u_j})^2 \right).$$

The estimator is then defined as

$$S^u_{N,\mathrm{Cl}} = \frac{\frac{1}{N}\sum Y_i Y_i^{u_1} - \left(\frac{1}{2N}\sum (Y_i + Y_i^{u_1}) \right)^2}{\frac{1}{N}\sum M^u_i - \left(\frac{1}{N}\sum Z^u_i \right)^2}. \tag{A.7}$$

This estimator was first introduced by Monod in (37) and Janon et al. studied its asymptotic properties (CLT, efficiency) in (31). In (41, 40) Owen introduces new estimators for Sobol indices and compares numerically their performances. The delta method can also be used on these pick-freeze estimators to derive their asymptotic properties.

Remark: One could use all the information available in the sample by defining the following estimator:

$$\frac{\frac{1}{N}\sum Y_i Y_i^{u_1} - \left(\frac{1}{N}\sum Z_i^u\right)^2}{\frac{1}{N}\sum M_i^u - \left(\frac{1}{N}\sum Z_i^u\right)^2}.$$

However, our empirical studies show that this estimator has a larger variance than $T_{N,\mathrm{Cl}}^u$.

Asymptotic results: consistency and asymptotic normality

The estimator $S_{N,\mathrm{Cl}}^u$ is consistent in estimating S_{Cl}^u. Moreover if we assume that $\mathbb{E}(Y^4) < \infty$, then

$$\sqrt{N}\left(S_{N,\mathrm{Cl}}^u - S_{\mathrm{Cl}}^u\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \sigma^2\right) \tag{A.8}$$

where

$$\sigma^2 = \frac{\mathrm{Var}(YY^u) - 2S_{\mathrm{Cl}}^u \mathrm{Cov}(YY^u, M^u) + (S_{\mathrm{Cl}}^u)^2 \mathrm{Var}(M^u)}{\left(\mathrm{Var}(Y)\right)^2}.$$

One can generalize this estimation procedure and asymptotic results to subset u of $I_p = \{1, \cdots, p\}$ (instead of singletons) and to vectors of such indices. Then one can construct confidence regions and build tests of significance.

A.2 A generalization of the Sobol indices for multivariate outputs

In this subsection, we generalize the procedure of the previous one to multivariate outputs: the output Y is now given by

$$Y = f(X_1, \ldots, X_p) = \begin{pmatrix} f_1(X_1, \ldots, X_p) \\ \vdots \\ f_k(X_1, \ldots, X_p) \end{pmatrix},$$

where $f : E \to \mathbb{R}^k$ is still an unknown deterministic measurable function. We assume that $X_1, \ldots, X_p$ are independent and that Y is square integrable (i.e. $\mathbb{E}\left(\|Y\|^2\right) < \infty$). We also assume, without loss of generality, that the covariance matrix of Y is positive definite.

A.2.1 Motivation

One can easily be convinced that neither $(S_{\mathrm{Cl}}^u(f_1), \cdots, S_{\mathrm{Cl}}^u(f_k))$ nor $S_{\mathrm{Cl}}^u(\|Y\|^2)$ are suitable ways to define the sensitivity of the multivariate output Y. Another motivation to introduce new Sobol indices is related to the statistical problem of their estimation. As the dimension increases, the statistical estimation of the whole vector of scalar Sobol indices becomes more and more expensive. Moreover, the interpretation of such a large vector is not easy. This strengthens the fact that one needs to introduce Sobol indices of small dimension, which condense all the information contained in a large collection of scalars.

A.2.2 Definition of a new index for multivariate outputs

Proceeding as done previously and computing the covariance matrix of both sides of (A.2) leads to

$$\Sigma = C_u + C_{\sim u} + C_{u,\sim u}. \tag{A.9}$$

Here Σ, C_u, $C_{\sim u}$ and $C_{u,\sim u}$ are denoting respectively the covariance matrices of Y, $f_u(X_{\mathbf{u}})$, $f_{\sim u}(X_{\sim \mathbf{u}})$ and $f_{u,\sim u}(X_{\mathbf{u}}, X_{\sim \mathbf{u}})$.
In the general case ($k \geqslant 2$), the equation (A.9) can be scalarized in the following way

$$\mathrm{Tr}(\Sigma) = \mathrm{Tr}(C_{\mathbf{u}}) + \mathrm{Tr}(C_{\sim \mathbf{u}}) + \mathrm{Tr}(C_{\mathbf{u},\sim \mathbf{u}}).$$

This suggests to define as soon as $\mathrm{Tr}(\Sigma) \neq 0$ (which is the case as soon Y is not constant), the sensitivity measure of Y with respect to X_u as

$$S^u(f) = \frac{\mathrm{Tr}(C_{\mathbf{u}})}{\mathrm{Tr}(\Sigma)}.$$

Of course we can analogously define

$$S^{\sim u}(f) = \frac{\mathrm{Tr}(C_{\sim \mathbf{u}})}{\mathrm{Tr}(\Sigma)}, \quad S^{u,\sim u}(f) = \frac{\mathrm{Tr}(C_{\mathbf{u},\sim \mathbf{u}})}{\mathrm{Tr}(\Sigma)}.$$

The following straightforward properties are natural requirements for a sensitivity measure.

Properties

1. The generalized sensitivity measures sum up to 1

$$S^u(f) + S^{\sim u}(f) + S^{u,\sim u}(f) = 1. \tag{A.10}$$

2. $0 \leqslant S^u(f) \leqslant 1$.
3. $S^u(f)$ is invariant by left-composition of f by any isometry of $\mathbb{R}^k$ i.e.

   for any square matrix O of size k s.t. $O^t O = \mathrm{Id}_k$, $S^u(Of) = S^u(f)$;

4. $S^u(f)$ is invariant by left-composition by any nonzero scaling of f i.e.

   for any $\lambda \in \mathbb{R}$, $S^u(\lambda f) = S^u(f)$;

5. For $k = 1$, we recover the standard definition of the scalar Sobol index.

Variational formulation
We assume here that $\mathbb{E}(Y) = 0$, if it is not the case, one has to consider the centered variable $Y - \mathbb{E}(Y)$. As in dimension 1, one can see that this new index can also be seen as the solution of the following least-squares problem (see (31))

$$\underset{a}{\mathrm{Argmin}}\, \mathbb{E}\|Y^u - aY\|^2.$$

As a consequence, $S^u(f)Y$ can be seen as the projection of Y^u on $\{aY, a \in \mathbb{R}\}$.

Examples: *(1) We consider as first example*

$$Y = f^a(X_1, X_2) = \begin{pmatrix} aX_1 \\ X_2 \end{pmatrix},$$

with X_1 and X_2 i.i.d. standard Gaussian random variables. We easily get

$$S^{\mathbf{1}}(f^a) = \frac{a^2}{a^2+1} \quad \textit{and} \quad S^{\mathbf{2}}(f^a) = \frac{1}{a^2+1} = 1 - S^{\mathbf{1}}(f).$$

(2) We consider as second example

$$Y = f^{a,b}(X_1, X_2) = \begin{pmatrix} X_1 + X_1X_2 + X_2 \\ aX_1 + bX_1X_2 + X_2 \end{pmatrix}.$$

We have

$$S^{\mathbf{1}}(f^{a,b}) = \frac{1+a^2}{4+a^2+b^2} \quad \textit{and} \quad S^{\mathbf{2}}(f^{a,b}) = \frac{2}{4+a^2+b^2}$$

and obviously

$$S^{\mathbf{1}}(f^{a,b}) \geqslant S^{\mathbf{2}}(f^{a,b}) \iff a^2 \geqslant 1.$$

This result has the natural interpretation that, as X_1 is scaled by a, it has more influence if and only if this scaling enlarges X_1's support i.e. $|a| > 1$.

A.2.3 Estimation of $S^u(f)$ and asymptotic properties

In practice, the covariance matrices $C_{\mathbf{u}}$ and Σ are not analytically available. As seen previously, in the scalar case ($k = 1$), it is customary to estimate $S^u(f)$ by using a Monte-Carlo Pick and Freeze method (49, 31), which uses a finite sample of evaluations of f. In the multivariate context, we propose a Pick and Freeze estimator for the vectorial case which generalizes the estimator $S^u_{N,\mathrm{Cl}}$ studied in (31).

Design of experiments and estimation phase

Let N be an integer. We take N independent copies $Y_1, \ldots, Y_N$ (resp. $Y_1^u, \ldots, Y_N^u$) of Y (resp. Y^u). For $l = 1, \ldots, k$, and $i = 1, \ldots, N$, we also denote by $Y_{i,l}$ (resp. $Y_{i,l}^u$) the l^{th} component of Y_i (resp. Y_i^u). We then define the following estimator of $S^u(f)$

$$S_{u,N} = \frac{\sum_{l=1}^k \left(\frac{1}{N} \sum_{i=1}^N Y_{i,l} Y_{i,l}^u - \left(\frac{1}{N} \sum_{i=1}^N \frac{Y_{i,l}+Y_{i,l}^u}{2} \right)^2 \right)}{\sum_{l=1}^k \left(\frac{1}{N} \sum_{i=1}^N \frac{Y_{i,l}^2+(Y_{i,l}^u)^2}{2} - \left(\frac{1}{N} \sum_{i=1}^N \frac{Y_{i,l}+Y_{i,l}^u}{2} \right)^2 \right)} = \frac{\operatorname{Tr}(C_{u,N})}{\operatorname{Tr}(\Sigma_N)}. \tag{A.11}$$

where $C_{u,N}$ and Σ_N are the empirical estimators of $C_u = \operatorname{Cov}(Y, Y^u)$ and $\Sigma = \operatorname{Var}(Y)$ defined by

$$C_{u,N} = \frac{1}{N} \sum_{i=1}^N Y_i^u Y_i^t - \left(\frac{1}{N} \sum_{i=1}^N \frac{Y_i + Y_i^u}{2} \right) \left(\frac{1}{N} \sum_{i=1}^N \frac{Y_i + Y_i^u}{2} \right)^t$$

and

$$\Sigma_N = \frac{1}{N}\sum_{i=1}^{N} \frac{Y_i Y_i^t + Y_i^u (Y_i^u)^t}{2} - \left(\frac{1}{N}\sum_{i=1}^{N} \frac{Y_i + Y_i^u}{2}\right)\left(\frac{1}{N}\sum_{i=1}^{N} \frac{Y_i + Y_i^u}{2}\right)^t .$$

Roughly speaking, it amounts to consider the empirical estimations of the covariance and variance matrices and to take their traces.

Asymptotic results: consistency and asymptotic normality

Under some mild conditions, $S_{u,N}$ is consistent and asymptotically normal i.e. we have:

$$\sqrt{N}(S_{u,N} - S^{u,}(f)) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, \sigma^2)$$

where the limiting variance σ^2 can be computed explicitly.

Following the same idea, it is possible, for u and $v \subset \{1, \ldots, p\}$, to derive a (multivariate) central limit theorem for

$$(S_{u,N}, S_{v,N}, S_{u\cup v,N}) = \left(\frac{\mathrm{Tr}\,(C_{u,N})}{\mathrm{Tr}\,(\Sigma_N)}, \frac{\mathrm{Tr}\,(C_{v,N})}{\mathrm{Tr}\,(\Sigma_N)}, \frac{\mathrm{Tr}\,(C_{u\cup v,N})}{\mathrm{Tr}\,(\Sigma_N)}\right).$$

We then can derive a (scalar) central limit theorem for $S_{u\cup v,N} - S_{u,N} - S_{\mathbf{v},N}$, a natural estimator of $S^{u\cup v} - S^u - S^v$, which quantifies the influence (for $u \cap v = \emptyset$) of the interaction between the variables of u and v.

A.3 A generalization of Sobol indices for functional outputs

In many practical situations the output Y is functional. It is then useful to extend the vectorial indices to functional outputs.

A.3.1 Definition of the Sobol index

Let $\mathbb{H}$ be a separable Hilbert space endowed with the scalar product $\langle\cdot,\cdot\rangle$ and the norm $||\cdot||$. Let f be a $\mathbb{H}$-valued function, i.e. Y and Y^u are $\mathbb{H}$-valued random variables. We assume that $\mathbb{E}\left(\|Y\|^2\right) < \infty$. Recall that $\mathbb{E}\,(Y)$ is defined by duality as the unique member of $\mathbb{H}$ satisfying

$$\mathbb{E}\,(\langle h, Y\rangle) = \langle h, \mathbb{E}(Y)\rangle \quad \text{for all} \quad h \in \mathbb{H}.$$

From now on assume that Y is centered.

Recall that the covariance operator associated with Y is the endomorphism Γ on $\mathbb{H}$ defined, for $h \in \mathbb{H}$ by $\Gamma(h) = \mathbb{E}\,[\langle Y, h\rangle Y]$. We also recall that it is a well known fact that $\mathbb{E}\left(\|Y\|^2\right) < \infty$ implies that Γ is then a Trace class operator and its trace is then well defined. We generalize the definition of $S^u(f)$ introduced for multivariate outputs in the previous section to functional outputs.

Now using the so-called polar decomposition of the traces of Γ and Γ_u

$$\begin{aligned}\operatorname{Tr}(\Gamma) &= \mathbb{E}\left(\|Y\|^2\right) - \|\mathbb{E}(Y)\|^2 \\ \operatorname{Tr}(\Gamma_u) &= \frac{1}{4}\left[\mathbb{E}\left(\|Y+Y^u\|^2\right) - \mathbb{E}\left(\|Y-Y^u\|^2\right) - 4\|\mathbb{E}\left(Y\right)\|^2\right],\end{aligned}$$

we define the Sobol index of the functional output Y by

$$S^{u,\infty}(f) = \frac{\operatorname{Tr}(\Gamma_u)}{\operatorname{Tr}(\Gamma)} = \frac{1}{4}\frac{\mathbb{E}\left(\|Y+Y^u\|^2\right) - \mathbb{E}\left(\|Y-Y^u\|^2\right) - 4\|\mathbb{E}\left(Y\right)\|^2}{\mathbb{E}\left(\|Y\|^2\right) - \|\mathbb{E}(Y)\|^2},$$

where Γ_u is the endomorphism on $\mathbb{H}$ defined by $\Gamma_u(h) = \mathbb{E}\left[\langle Y^u, h\rangle Y\right]$ for any $h \in \mathbb{H}$.

A.3.2 Estimation of $S^{u,\infty}(f)$ and asymptotic properties

Let $(\varphi_l)_{1\leqslant l}$ be an orthonormal basis of $\mathbb{H}$. Then

$$\|Y\|^2 = \sum_{i=1}^{\infty}\langle Y, \varphi_i\rangle^2.$$

Now, in order to proceed to the estimation of $S^{u,\infty}(f)$ (and thus first to the estimation of $\operatorname{Tr}(\Gamma)$ and $\operatorname{Tr}(\Gamma_u)$), we truncate the previous sum by setting

$$\|Y\|_m^2 = \sum_{i=1}^{m}\langle Y, \varphi_i\rangle^2.$$

It amounts to truncate the expansion of Y to a certain level m. Let Y_m be the truncated approximation of Y:

$$Y_m = \sum_{l=1}^{m}\langle Y, \varphi_i\rangle\varphi_l,$$

seen as a vector of dimension m. Thus the results of Section A.2 can be applied to Y_m. Notice that Y_m is then the projection of Y onto $\operatorname{Span}(\varphi_1, \ldots, \varphi_m)$.

Design of experiments and estimation phase

Let N be an integer. As done in Section A.2, we take N independent copies $Y_1, \ldots, Y_N$ (resp. $Y_1^u, \ldots, Y_N^u$) of Y (resp. Y^u). We define the following estimator of $S^{u,\infty}(f)$:

$$S_{u,m,N} = \frac{\frac{1}{4N}\sum_{i=1}^{N}\left(\|Y_i+Y_i^u\|_m^2 - \|Y_i-Y_i^u\|_m^2 - \|\overline{Y}+\overline{Y^u}\|_m^2\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|Y_i\|_m^2+\|Y_i^u\|_m^2}{2} - \left\|\frac{\overline{Y}+\overline{Y^u}}{2}\right\|_m^2\right)}.$$

Asymptotic results: consistency and asymptotic normality

Under some mild conditions, $S_{u,m,N}$ is consistent and asymptotically normal i.e. we have:

$$\sqrt{N}(S_{u,m,N} - S^{u,\infty}(f)) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0,\sigma^2)$$

where the limiting variance σ^2 can be computed explicitly.

B A survey on moment independent measures

As defined in the previous sections, notice that the Sobol indices and their Monte Carlo estimation are based on order two methods since they derived from the L^2 Hoeffding functional decomposition. This is the main drawback of this kind of methods. As an illustration, imagine an output Y that is a symmetric function of the inputs X_1 and X_2 that do not share the same distribution but have the same first four moments and satisfy $\mathbb{E}\left[X_1^5\right] \neq \mathbb{E}\left[X_2^5\right]$. Necessarily, X_1 and X_2 should not have the same importance that will not be traduced on the Sobol indices.

Moreover, Sobol indices are based on L^2 decomposition; they are well adapted to measure the contribution of an input to the deviation around the mean of Y. However, it seems very intuitive that the sensitivity of an extreme quantile of Y could depend on sets of variables different from that highlighted when studying the variance sensitivity. Thus the same index should not be used for these two different tasks and we need to define more adapted indices.

Hence the need to introduce a new sensitivity index that takes into account such an importance. There are several ways to generalize the Sobol indices. One can, for example, define new indices through contrast functions based on the quantity of interest (see (18)). Unfortunately the Monte Carlo estimator of these new indices are computationally expensive. Now, as pointed out in (38), (39), (2), (7) and (6) there are situations where higher order methods give a sharper analysis on the relative influence of the input and allow finer screening procedures. Borgonovo et al. propose and study an index based on the total variation distance (see (2), (7) and (6)). While Owen et al. suggest to use procedures based on higher moments (see (38), (39)). The first index presented in the following section follows these tracks.

B.1 Multiple Pick and Freeze method

Following (38) and (39), we generalize the numerator of the classical Sobol index quantity by considering higher order moments. Indeed, for $p \geqslant 3$, set

$$H_v^p := \mathbb{E}\left[(\mathbb{E}[Y|X_v] - \mathbb{E}[Y])^p\right] \tag{B.1}$$

in such a way that the numerator of the scalar Sobol index is H_v^2. The following result gives the Pick and Freeze representation of H_v^p:

$$\mathbb{E}\left[(\mathbb{E}[Y|X_v] - \mathbb{E}[Y])^p\right] = \mathbb{E}\left[\prod_{i=1}^{p}\left(Y^{v,i} - \mathbb{E}[Y]\right)\right]. \tag{B.2}$$

Here, $Y^{v,1} = Y$ and for $i = 2, \dots, p$, $Y^{v,i}$ is constructed independently as Y^v.

Remarks

1. H_v^p is only non negative for even p.
2. For any p,

$$|H_v^p| \leqslant \mathbb{E}\left[|Y - \mathbb{E}[Y]|^p\right].$$

3. As the classical Sobol index, H_v^p is still invariant by translation of the output.

Design of experiments and estimation phase

In view of the estimation of H_v^p, we first develop the product in the right-hand side of (B.2) to get that

$$H_v^p = \sum_{l=0}^{p} \binom{p}{l} (-1)^{p-l} \mathbb{E}\left[Y\right]^{p-l} \mathbb{E}\left[\prod_{i=1}^{l} Y^{v,i}\right].$$

with the usual convention $\prod_{i=1}^{0} Y^{v,i} = 1$. Second we use a Monte Carlo scheme and consider the following Pick and Freeze design of experiment constituted by the following $p \times N$-sample

$$\left(Y_j^{v,i}\right)_{(i,j)\in I_p \times I_N}.$$

We define for any any $N \in \mathbb{N}^*$, $j \in I_N$ and $l \in I_p$,

$$P_{l,j}^v = \binom{p}{l}^{-1} \sum_{k_1 < \ldots < k_l \in I_p} \left(\prod_{i=1}^{l} Y_j^{v,k_i}\right) \quad \text{and} \quad \overline{P}_l^v = \frac{1}{N}\sum_{j=1}^{N} P_{l,j}^v.$$

The Monte Carlo estimator is then

$$H_{p,N}^v = \sum_{l=0}^{p} \binom{p}{l} (-1)^{p-l} \left(\overline{P}_1^v\right)^{p-l} \overline{P}_l^v. \tag{B.3}$$

Asymptotic results: consistency and asymptotic normality

$H_{p,N}^v$ is consistent and asymptotically normal:

$$\sqrt{N}\left(H_{p,N}^v - H_p^v\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \sigma^2\right) \tag{B.4}$$

where σ^2 can be explicitly computed.

The collection of all indices H_v^p is much more informative than the classical Sobol index. Nevertheless it has several drawbacks: it may be negative when p is odd. To overcome this fact, we may have introduced $\mathbb{E}\left[|\mathbb{E}[Y|X_i, i \in v] - \mathbb{E}[Y]|^p\right]$ but proceeding in such a way, we would have loose the Pick and Freeze estimation procedure. The Pick and Freeze estimation procedure is computationally expensive: it requires a $p \times N$ sample of the output Y. In a sense, if we want to have a good idea of the influence of an input on the law of the output, we need to estimate the first d indices H_v^p and hence we need to run the black-box code $K \times N$ times. In the next section, we introduce a new sensitivity index that is based in the conditional distribution of the output and requires only $3 \times N$.

B.2 The Cramér von Mises index

In this section, the code will be denoted by $Z = f(X_1, \ldots, X_d) \in \mathbb{R}$. Let F be the distribution function of Z

$$F(t) = \mathbb{P}\left(Z \leqslant t\right) = \mathbb{E}\left[\mathbb{1}_{\{Z \leqslant t\}}\right]$$

and $F^v(t)$ the conditional distribution function of Z conditionally on X_v:

$$F^v(t) = \mathbb{P}\left(Z \leqslant t | X_v,\right) = \mathbb{E}\left[\mathbb{1}_{\{Z \leqslant t\}} | X_v\right].$$

It is obvious that $\mathbb{E}\left[F^v(t)\right] = F(t)$. We apply the framework presented in Section B.1 with $Y(t) = \mathbb{1}_{\{Z \leqslant t\}}$ and $p = 2$. We then have a consistent and asymptotically normal estimation procedure for the estimation of

$$\mathbb{E}\left[(F(t) - F^v(t))^2\right].$$

We define a Cramér Von Mises type distance of order 2 between $\mu := \mathcal{L}(Z)$ and $\mathcal{L}(Z|X_v)$ by

$$D^v_{2,CVM} := \int_{\mathbb{R}} \mathbb{E}\left[(F(t) - F^v(t))^2\right] d\mu(t). \tag{B.5}$$

The aim of the rest of the section is dedicated to the estimation of $D^v_{2,CVM}$ and the study of the asymptotic properties of the estimator. Notice that

$$D^v_{2,CVM} = \mathbb{E}\left[\mathbb{E}\left[(F(Z) - F^v(Z))^2\right]\right]. \tag{B.6}$$

Properties

1. $0 \leqslant D^v_{2,CVM} \leqslant \dfrac{1}{4}$. Moreover, if F is continuous, we have $0 \leqslant D^v_{2,CVM} \leqslant \dfrac{1}{6}$.
2. As the classical Sobol index, $D^v_{2,CVM}$ is still invariant by translation, by left-composition by any nonzero scaling of Y and by left-composition of Y by any isometry.

Design of experiments and estimation phase

We then proceed to a double Monte-Carlo scheme for the estimation of $D^v_{2,CVM}$ and consider the following design of experiments consisting in:

1. two N-samples of Z: $(Z_j^{v,1}, Z_j^{v,2})$, $1 \leqslant j \leqslant N$;
2. a third N-sample of Z independent of $(Z_j^{v,1}, Z_j^{v,2})_{1 \leqslant j \leqslant N}$: W_k, $1 \leqslant k \leqslant N$.

The natural estimator of $D^v_{2,CVM}$ is then given by

$$\widehat{D}^v_{2,CVM} = \frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}_{\{Z_j^{v,1} \leqslant W_k\}}\mathbb{1}_{\{Z_j^{v,2} \leqslant W_k\}} - \left[\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\{Z_j^{v,1} \leqslant W_k\}} + \mathbb{1}_{\{Z_j^{v,2} \leqslant W_k\}}\right)\right]^2\right\}.$$

Asymptotic results: consistency and asymptotic normality

The estimator $\widehat{D}^v_{2,CVM}$ is consistent as N goes to infinity. Moreover, the sequence of estimators $\widehat{D}^v_{2,CVM}$ is asymptotically normal in estimating $D^v_{2,CVM}$ that is $\sqrt{N}\left(\widehat{D}^v_{2,CVM} - D^v_{2,CVM}\right)$ is weakly convergent to a Gaussian centered variable whose variance can be explicitly computed.

B.3 Other moment independent measures

Alternative definitions for measuring the strength of the statistical dependence of Y on X_k have been proposed, giving rise to the class of distribution-based sensitivity measures. They define the importance of X_k as the distance between the unconditional distribution of Y and its conditional distribution (see (1, 4, 5) for reviews). These sensitivity measures are defined both in the presence and absence of correlations. Three examples of such sensitivity measures are

- **u** the δ importance measure based on the L^1-norm between densities (3):

  $$\delta_k = \frac{1}{2}\mathbb{E}\left[\int |p_Y(y) - p_{Y|X_k}(y)|\, dy\right]$$

  where $p_Y(y)$ and $p_{Y|X_k}(y)$ stands respectively for the density function of Y and $Y|X_k$. Notice that $\delta_k = 0$ if and only if Y is independent of X_k.

- **u** the β^{KS} sensitivity measure based on the Kolmogorov-Smirnov separation between cumulative distributions functions:

  $$\beta_k = \mathbb{E}\left[\sup_y |F_Y(y) - F_{Y|X_k}(y)|\right].$$

  Both sensitivity measures δ_k and β_k are monotonic transformation invariant.

- **u** the θ probabilistic sensitivity measure based on the family of Shannon's cross-entropy:

  $$\theta_k = \mathbb{E}\left[\int |p_{Y|X_k}(y)(\log p_{Y|X_k}(y) - \log p_{Y|X_k}(y))|\, dy\right].$$

  θ_k can be interpreted as value of information sensitivity measures as the classical Sobol index (see e.g. (4) for more details). We will see in the next subsection that this measure is part of a larger lass of sensitivity measures based on dissimilarity distances and the family of Csiszár's divergences.

A common rationale

Variance-based and the previous distribution-based sensitivity measures have a common conceptual aspect. Using a given operator, they assess the change in decision-maker degree of belief about Y after she has come to know $X_k = x_k$. A closely-related field in which analysts are interested in computing the distance between distributions is information theory (Csiszár 2008). In information theory, the distributions are statistical signals. In probabilistic sensitivity analysis, they are the conditional and unconditional model output distributions, $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_k=x_k}$. We call

$$\xi_k = \mathbb{E}[\gamma_k(X_k)] = \mathbb{E}[\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_k})] \tag{B.7}$$

the *global sensitivity measure* of X_k based on operator $\zeta(\cdot,\cdot)$ and $\gamma_k(x_k)$ the *inner statistic* of ξ_k.

The above framework accommodates the definitions of the probabilistic sensitivity measures described previously. Selecting as inner operators

- **u** $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_k}) = \mathbb{E}[(Y - \mathbb{E}[Y])^2|X_k = x_k] = \mathbb{E}[(Y - \mathbb{E}[Y|X_k])^2|X_k = x_k]$;

u or $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_k}) = \dfrac{1}{2}\displaystyle\int |p_Y(y) - p_{Y|X_k}(y)|\, dy$;

u or $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_k}) = \sup\limits_y |F_Y(y) - F_{Y|X_k}(y)|$;

u or $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_k}) = \displaystyle\int |p_{Y|X_k}(y)(\log p_{Y|X_k}(y) - \log p_{Y|X_k}(y))|\, dy$;

we obtain the inner statistics of the classical Sobol index, δ_k, β_k and θ_k respectively. The properties and estimators of these sensitivity measures are presented in details in (4).

B.4 Moment independent measures and dissimilarity distances

This section is an extract of the synthesis of Da Veiga in (10) and presents another way to define moment independent measures through dissimilarity distances and encompasses some already known sensitivity indices.

Motivation

As pointed out by (1), a natural way of defining the impact of a given input X_k on Y is to consider a function which measures the similarity between the distribution of Y and that of $Y|X_k$. More precisely, the impact of X_k on Y is given by

$$S_{X_k} = \mathbb{E}_{X_k}[d(Y, Y|X_k)] \tag{B.8}$$

where $d(\cdot,\cdot)$ denotes a dissimilarity measure between two random variables. The advantage of such a formulation is that many choices for d are available and we will see in what follows that some natural dissimilarity measures yield sensitivity indices related to well known quantities. However before going further, let us note that the naïve dissimilarity measure

$$d(Y, Y|X_k) = (E[Y] - E[Y|X_k])^2 \tag{B.9}$$

where random variables are compared only through their mean values produces the unnormalized Sobol first-order sensitivity index $S^1_{X_k} = \mathrm{Var}(\mathbb{E}[Y|X_k])$.

Dissimilarity measures

Assuming all input random variables have an absolutely continuous distribution with respect to the Lebesgue measure on $\mathbb{R}$, the f-divergence between Y and $Y|X_k$ is given by

$$d_h(Y||Y|X_k) = \int_{\mathbb{R}} h\left(\frac{p_Y(y)}{p_{Y|X_k}(y)}\right) p_{Y|X_k}(y)\, dy$$

where h is a convex function such that $h(1) = 0$ and p_Y and $p_{Y|X_k}$ are the probability distribution functions of Y and $Y|X_k$, respectively. Standard choices for the function h include for example

u Kullback-Leibler divergence: $h(t) = -\ln(t)$ or $h(t) = t\ln(t)$;

u Hellinger distance: $h(t) = (\sqrt{t} - 1)^2$;

u Total variation distance: $h(t) = |t - 1|$;

u Pearson χ^2 divergence: $h(t) = (t - 1)^2$ or $h(t) = t^2 - 1$;

u Neyman χ^2 divergence: $h(t) = (t-1)^2/t$ or $h(t) = (1-t^2)/t$.

From dissimilarity measures to sensitivity indices

Plugging this dissimilarity measure in (B.8) yields the following sensitivity index:

$$S^h_{X_k} = \int_{\mathbb{R}^2} h\left(\frac{p_Y(y)p_{X_k}(x)}{p_{X_k,Y}(x,y)}\right) p_{X_k,Y}(x,y)\,dxdy \tag{B.10}$$

where p_{X_k} and $p_{X_k,Y}$ are the probability distribution functions of X_k and (X_k, Y), respectively. First of all, note that inequalities on Csiszár f-divergences imply that such sensitivity indices are positive and equal zero when Y and X_k are independent. Also, it is important to note that given the form of $S^h_{X_k}$, it is invariant under any smooth and uniquely invertible transformation of the variables X_k and Y. This is a major advantage over variance-based Sobol sensitivity indices, which are only invariant under linear transformations. It is easy to see that

u The total variation distance with $h(t) = |t-1|$ gives a sensitivity index equal to the one proposed by (3):

$$S^h_{X_k} = \int_{\mathbb{R}^2} |p_Y(y)p_{X_k}(x) - p_{X_k,Y}(x,y)|\,dxdy.$$

u The Kullback-Leibler divergence with $h(t) = -\ln(t)$ yields

$$S^h_{X_k} = \int_{\mathbb{R}^2} p_{X_k,Y}(x,y) \ln\left(\frac{p_{X_k,Y}(x,y)}{p_Y(y)p_{X_k}(x)}\right) dxdy,$$

that is the mutual information $I(X_k; Y)$ between X_k and Y. A normalized version of this sensitivity index was studied by (35).

u The Neyman χ^2 divergence with $h(t) = (1-t^2)/t$ leads to

$$S^h_{X_k} = \int_{\mathbb{R}^2} p_{X_k,Y}(x,y) \ln\left(\frac{p_{X_k,Y}(x,y)}{p_Y(y)p_{X_k}(x)}\right) dxdy,$$

which is the so-called squared-loss mutual information between X_k and Y (or mean square contingency).

These results show that some previously proposed sensitivity indices are actually special cases of more general indices defined through Csiszár f-divergences. To the best of our knowledge, this is the first work in which this link is highlighted. Moreover, the specific structure of equation (B.10) makes it possible to envision more efficient tools for the estimation of these sensitivity indices. Indeed, it only involves approximating a density ratio rather than full densities. But more importantly, we see that special choices for f define sensitivity indices that are actually well-known dependence measures such as the mutual information. This paves the way for completely new sensitivity indices based on recent state-of-the-art dependence measures.

Estimation phase

Coming back to equation (B.10), the goal is to estimate

$$S^h_{X_k} = \int_{\mathbb{R}^2} h\left(\frac{1}{r(x,y)}\right) p_{X_k,Y}(x,y)\,dxdy = \mathbb{E}_{(X_k,Y)}\left[h\left(\frac{1}{r(X_k,Y)}\right)\right]$$

where $r(x,y) = p_{X_k,Y}(x,y)/(p_Y(y)p_{X_k}(x))$ is the ratio between the joint density of (X_k, Y) and the marginals. Of course, straightforward estimation is possible if one estimates the densities $p_{X_k,Y}(x,y)$, $p_{X_k}(x)$ and $p_Y(y)$ with e.g. kernel density estimators. However, it is well known that density estimation suffers from the curse of dimensionality. This limits the possible multivariate extensions we discuss in the next subsection. Besides, since only the ratio function $r(x,y)$ is needed, we expect more robust estimates by focusing only on it.

Let us assume now that we have a sample $(X_{k,i}, Y_i)$ for $i = 1, ..., N$ of (X_k, Y), the idea is to build first an estimate $\hat{r}(x,y)$ of the ratio. The final estimator $\hat{S}^h_{X_k}$ of $S^h_{X_k}$ will then be given by

$$\hat{S}^h_{X_k} = \frac{1}{N}\sum_{i=1}^{N} h\left(\frac{1}{\hat{r}(X_{k,i}, Y_i)}\right).$$

Powerful estimating methods for ratios include e.g. maximum-likelihood estimation (51), unconstrained least-squares importance fitting (32), among others (50). A k-nearest neighbors strategy dedicated to mutual information is also discussed in(34).

C A survey on functional data clustering

We follow (29) and (16) and references therein. Functional data analysis extends the classical multivariate methods when data are functions or curves such as the evolution of some stock-exchange index, the size of an individual... A functional random variable X is a random variable with values in an infinite dimensional space. Then functional data represents a set of observation $\{X_1, \ldots, X_n\}$ of X. The underlying model for X_i's is generally an i.i.d. sample of random variables drawn from the same distribution X.

In modern statistical terminology, the word “classification” is used with two principal meanings: while “unsupervised classification” roughly stands for “clustering”, “supervised classification” is used as a synonym for the more classical name “discriminant analysis”. Both meanings correspond with those given by the Oxford English Dictionary under the entry “classify”: first, *arrange a group in classes according to shared characteristics*; second, *assign to a particular class or category.*

The basic aim of unsupervised classification techniques is to partition a sample $X_1, \ldots, X_n$ with a large n into a number k of clusters. These clusters are defined in such a way that the members of each class are “similar” to each other in a sense that is prescribed by the clustering algorithm used. The number k of clusters has to be given in advance or determined simultaneously by the clustering algorithm.

The supervised classification corresponds to the second meaning of “classify” mentioned above. Here k populations P_1, …, P_k are given and clearly defined in advance. The practitioner has a training sample $(X_i, Y_i)_{i=1,\ldots,n}$ where the X_i's are independent copies of a random variable X and each Y_i is the index of the corresponding population to which X_i belongs to. The distribution of X is assumed to be different in each population. The term “supervised” accounts for the fact that the elements of the training sample are supposed to be classified with no error. The goal is then to assign a new realization of X into one of the populations P_j.

C.1 Supervised classification

In order t fix ideas, we here concentrate on the binary problem: we assume that there is only two underlying populations P_0 and P_1. The available information is given by a training sample $(X_i, Y_i)_{i=1\dots n}$ where X_i is a n-sample of the functional feature $X \in \mathbb{H}$ and Y_i is the corresponding label ($Y_i = 0$ if X_i belongs to population 0 and $Y_i = 1$ otherwise).
The problem is to find a classifier $g : \mathbb{H} \to \{0, 1\}$ that minimizes the criterion error $\mathbb{P}(g(X) \neq Y)$. An optimal (but unknown) classifier is

$$g^\star(x) = \mathbb{1}_{\{\eta(x) \geqslant 1/2\}} \tag{C.1}$$

where $\eta(x) = \mathbb{E}(Y|X = x)$. In practice, we use the training sample to construct a good classifier:

$$g^\star(x) = \operatorname*{Argmin}_g \hat{L}_n(g)$$

that minimizes the empirical risk $\hat{L}_n(g) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n \mathbb{1}_{\{g(X_i)\neq Y_i\}}$.

This methodology is more efficient than the one that consists to estimate the regression function η.

Linear discrimination rules

Here the principle is to generalize the Fisher's methodology introduced in the multivariate case and consists in projecting the infinite dimensional x onto the real line and comparing such a projection with those of the mean functions $\mu_j(t) = \mathbb{E}(X(t)|Y = j)$, $j = 0, 1$. The projection "direction" β would be selected as the maximizer of the distance between the projected class means $\langle\beta, \mu_0\rangle$ and $\langle\beta, \mu_1\rangle$ which leads to maximizing

$$\frac{\mathrm{Var}(\mathbb{E}(\langle\beta, X\rangle|Y))}{\mathbb{E}(\mathrm{Var}(\langle\beta, X\rangle|Y))}$$

under the constraint $\displaystyle\int \beta(t)\langle\beta, \mathrm{Cov}(X(t), X(\cdot)|Y = j)\rangle dt = 1$, $j = 0, 1$. The problem is that the covariance operator associated with the kernel $\mathrm{Cov}(X(t), X(s)|Y = j)$ is not in general invertible and thus the above optimization problem has no solution that forces the practitioner to find approximated solutions.

k-NN rules

This technique, adapted in both the multivariate and the functional settings, consists in assigning the data x to the population P_0 as soon as the majority of the k-nearest neighbors of x belongs to P_0. It amounts to replace the unknown regression function η defined in (C.1) with the regression estimator

$$\eta_n(x) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}_{\{X_i \in k(x)\}} Y_i$$

where "$X_i \in k(x)$" means that X_i is one of the nearest neighbors of x. In practice the value of k can be chosen by minimizing

$$\tilde{L}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{g_{ni}(X_i)\neq Y_i\}}$$

where g_{ni} denotes the leave-one-out k-NN rule based on the original sample of size n in which the ith observation (X_i, Y_i) has been deleted.

Kernel rules

In the same spirit of the kNN rule there exists another simple classification procedure, the moving window rule, which is based on the majority vote of the data surrounding the point x to be classified that assigns x to P_0 if

$$\sum_{i=1}^{n} \mathbb{1}_{\{Y_i=0,\ D(X_i,x)\leqslant h\}} > \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=1,\ D(X_i,x)\leqslant h\}}$$

and to P_1 otherwise. A smoother and more general version is given by

$$g_n(x) = \begin{cases} 0 & \text{if} \quad \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=0\}} K\left(\frac{D(X_i,x)}{h}\right) > \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=1\}} K\left(\frac{D(X_i,x)}{h}\right) \\ 1 & \text{otherwise} \end{cases}$$

where the Kernel is a non increasing known function. Some popular choices for K are the Gaussian kernel $K(x) = e^{-x^2}$, the Epanechnikov kernel $K(x) = (1-x^2)\mathbb{1}_{[0,1]}(x)$ or even the uniform kernel $K(x) = \mathbb{1}_{[0,1]}(x)$. Notice that it amounts to replace the unknown regression function η defined in (C.1) with the kernel regression estimator

$$\eta_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{D(X_i,x)}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{D(X_i,x)}{h}\right)}$$

Classification based on partial least squares (PLS)

The general idea of PLS classification is to reduce the dimension of the data via PLS and then to use any standard classification method (Fisher's linear procedure for example) on the projected data of lower dimension.

The idea behind PLS is similar to that of principal components: to project the data along directions of high variability. The main difference is that PLS also takes into account the variable response Y when defining the projection directions. More precisely, let X be a p-dimensional random vector of explanatory variables with covariance matrix Σ_X and Y the real random response with variance σ^2. Denote Σ_{XY} the $p \times 1$ matrix of covariances between X and Y. The first pair of PLS directions are the unit vector a_1 and the scalar b_1 that maximizes with respect to a and b the expression

$$\frac{(\mathrm{Cov}(a^T X, bY))^2}{b^2(a^T a)};$$

in fact, a_1 is the eigenvector of $\Sigma_{XY}\Sigma_{YX}$ corresponding to the largest eigenvalue of this matrix and b_1 fulfills $b_1 = \Sigma_{XY} a_1$. The remaining PLS directions can be found in a similar way by imposing the additional condition that the next direction a_{k+1} is orthogonal to the previous ones. In our particular binary context, the matrix $\Sigma_{XY}\Sigma_{YX}$ can be naturally estimated by

$$S_{XY}S_{YX} = \sum_{i=0}^{1} \frac{1}{(n-1)^2} n_i^2 (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

where $\bar{x} = (n_0\bar{x}_0 + n_1\bar{x}_1)/(n_0+n_1)$ and for $i = 0, 1$, $\bar{x}_i$ denotes the vector of means estimated from the n_i sample elements belonging to the population P_i.

Classification based on reproducing kernels

In this methodology, we use a plug-in classifier obtained by replacing the regression function η in (C.1) with the function $\hat{\eta} \in \mathcal{H}_k$ (the RKHS associated with the kernel k) minimizing the "regularized empirical risk"

$$\frac{1}{n}\sum_{i=1}^{n} C(X_i, Y_i, \hat{\eta}(X_i)) + J(\hat{\eta})$$

where C is a loss function convex to the third argument and J is a penalty term. One could take $C(X, Y, \hat{\eta}(X)) = (Y - \hat{\eta}(X))^2$ and $J(\hat{\eta}) = \lambda \|\hat{\eta}\|_k^2$ or even in the binary case $C(X, Y, \hat{\eta}(X)) = -Y \log \hat{\eta}(X) + \log(1 - 1/(1 - \hat{\eta}(X))$ the logistic loss function.

Classification based on depth measures

Assume we have a measurement $D(P_i, x)$ of a data x in the population P_i, $i = 0, 1$. For example, in the real line, the depth of a data x with respect to a population P defined through its distribution function F can be $F(x)(1 - F(x-))$ or even $\min\{F(x), 1 - F(x)\}$. Assume also we have empirical versions $D_{ni}(x)$ of these depth measures. A natural classifier would be $g(x) = \mathbb{1}_{\{D_{n1}(x) > D_{n0}(x)\}}$ which amounts to assign x to the population in which we estimate it is most deeply placed.

Other techniques of classification have been developed and rely on support vector machine (44) or even on neural networks (17).

C.2 Unsupervised classification

As mentioned before, cluster analysis deals with the problem of identifying groups, relatively isolated form each other, of similar points. Unlike supervised classification, there is no training sample to serve as a guide. In the finite dimensional case, there are two main techniques for cluster analysis: k-means and hierarchical clustering. The first one is an extension of the notion of mean. The second one relies on the use of a matrix of distance between the sample points. This two methods are presented in the sequel.

Classically, as we will explained later, when dealing with functional data, we will adopt two principal strategies that consists in:

1. adapting the k-means technique or the hierarchical clustering one to the functional case choosing dissimilarity distances adapted to this case,
2. or reducing the dimension and using the clustering techniques of the multivariate case.

In the next section, we present the functional principal component analysis, the k-means technique and the hierarchical clustering one in the multivariate case.

C.2.1 Some preliminaries

Functional principal component analysis

From the set of functional data $\{X_1, \ldots, X_n\}$, one is interested in an optimal representation of curves into a function space of reduced (finite) dimension. Functional principal component analysis (FPCA) has been introduced to address this problem and is widely used in data clustering. In order to fix ideas, we assume in this paragraph that X lives in L^2 and moreover that X is centered and L^2-continuous (which is not really restrictive):

$$\forall t \in \mathcal{T},\ \lim_{h\to 0} \mathbb{E}[|X_{t+h} - X_t|^2] = 0.$$

As done in Section A.3, the covariance operator associated with X is the endomorphism Γ on L_2 defined by, for $h \in L^2$,

$$\Gamma(h) = \mathbb{E}\left[\langle X, h\rangle X\right] = \int_0^T \gamma(\cdot, t) h(t)\, dt$$

where the kernel γ is defined by $\gamma(s,t) = \mathbb{E}[X_s X_t]$, $s,\ t \in \mathcal{T}$. The spectral analysis of Γ provides a countable set of positive eigenvalues $(\lambda_j)_{j\geqslant 1}$ associated to an orthonormal basis of eigenfunctions $(f_j)_{j\geqslant 1}$:

$$\Gamma(f_j) = \lambda_j f_j, \tag{C.2}$$

with $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$ and $\int_0^T f_j(t) f_{j'}(t)\, dt = 1$ if $j \neq j'$ and 0 otherwise. The principal components $(C_j)_{j\geqslant 1}$ of X are random variables defined as the projection of X on the eigenfunctions of Γ:

$$C_j = \int_0^T \gamma(\cdot, t) X_t f_j(t)\, dt = \langle X, f_j\rangle.$$

The principal components $(C_j)_{j\geqslant 1}$ are zero-mean uncorrelated random variables with variance λ_j, $j \geqslant 1$. With these definitions, the Karhunen-Loeve expansion holds:

$$X(t) = \sum_{j\geqslant 1} C_j f_j(t),\ t \in \mathcal{T}.$$

Truncating at the first q terms, one obtains the best approximation in norm L_2 of X_t by a sum of quasi-deterministic processes:

$$X(t)^{(q)} = \sum_{j=1}^{q} C_j f_j(t),\ t \in \mathcal{T}.$$

Computational methods for FPCA: As explained previously, the functional data are generally observed at discrete time points and a common solution to reconstruct the functional data is to assume that they belong to a finite dimensional space spanned by some basis of functions $\Phi = \{\phi_1, \ldots, \phi_L\}$:

$$X_i(t) = \Phi(t)'\alpha_i, \text{ with } \quad \alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,L}).$$

The eigenfunctions f_j belonging to the linear space spanned by the basis Φ, one may rewrite

$$f_j(t) = \Phi(t)'\beta_j, \text{ with } \quad \beta_j = (\beta_{i,1}, \ldots, \beta_{i,L}).$$

We then use the classical empirical estimation of the covariance kernel γ and solve

$$\hat{\Gamma}(\Phi(t)'\beta_j) = \lambda_j \Phi(t)'\beta_j,$$

instead of (C.2). The problem to solve is now linear.

k-means in the multivariate case

Very popular for cluster analysis in data mining, k-means clustering is a method of vector quantization, originally introduced in signal processing. Concretely, given a set of observations $(x_1, x_2, \ldots, x_n)$, k-means clustering aims to partition the n observations into $k \leqslant n$ sets $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares. In other words, its objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{m}_i\|^2$$

where m_i is the mean of points in S_i. This results in a partitioning of the data space into Voronoi cells. k-means clustering use cluster centers to model the data, like the expectation-maximization algorithm for mixtures of Gaussian distributions but tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

Given an initial set of k-means $m_1^{(1)}, \ldots, m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps:

1. **Assignment step**: assign each observation to the cluster whose mean yields the least within-cluster sum of squares. Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.

   $$S_i^{(t)} = \big\{x_p : \big\|x_p - m_i^{(t)}\big\|^2 \le \big\|x_p - m_j^{(t)}\big\|^2 \ \forall j, 1 \le j \le k\big\},$$

   where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

2. **Update step**: calculate the new means to be the centroids of the observations in the new clusters.

   $$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j.$$

   Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares objective.

The algorithm has converged when the assignments no longer change. Since both steps optimize the objective, and there only exists a finite number of such partitioning, the algorithm must converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm.

Hierarchical clustering in the multivariate case

In data mining, hierarchical clustering is a method of cluster analysis that builds a hierarchy of clusters, generally according to two techniques:

1. agglomerative which is a "botton up" approach where each observation starts in its own cluster and pairs of clusters are merged as one moves up iteratively the hierarchy.

2. divisive wich is a "top down" approach where all the observations start in the same initial cluster and splits are performed recursively as one moves down the hierarchy.

The results are usually presented in a dendogram.

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required; this is achieved by the use of

1. an appropriate metric (a measure of distance between pairs of observations). Usually, one uses the Euclidian distance or its square, the L^1 or L^∞ distances;
2. a linkage criterion that specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Some commonly used linkage criteria between two sets of observations A and B are:
   - **u** the maximum or complete linkage clustering: $\max\{d(a,b)/\ a \in A,\ b \in B\}$,
   - **u** the minimum or single-linkage clustering: $\min\{d(a,b)/\ a \in A,\ b \in B\}$,
   - **u** the mean or average linkage clustering: $\displaystyle\frac{1}{|A||B|}\sum_{a\in A,\ b\in B} d(a,b)$,
   - **u** the centroid linkage clustering: $\|c_s - c_t\|$ where c_s and $c - t$ are the centers of clusters s and t respectively.

C.2.2 Major functional data clustering approaches

Two-stages approaches

These approaches consist of a first step in which the dimension of the data is reduced and of a second step in which classical clustering tools for finite dimensional data are used: like e.g. k-means or hierarchical clustering (see Section C.2.1). The reducing dimension step consists generally in approximating the curves into a finite basis of functions: using splines or functional principal component analysis (see Section C.2.1).

Nonparametric approaches

Nonparametric approaches for functional data clustering are divided into two categories: methods who apply usual nonparametric clustering techniques with specific distances or dissimilarities (k-means or hierarchical clustering for functional data) and methods which propose new heuristics or geometry criteria to cluster functional data.

In the first category of methods, the proximity of two curves x and y is measured through

$$d_l(x,y) = \left(\int_{\mathcal{T}} (x^{(l)}(t) - y^{(l)}(t))^2 dt\right)^{1/2}$$

where $x^{(l)}$ is the l-th derivative of x. Usually the proximity measures d_0, d_1 or even $(d_0^2 + d_1^2)^{1/2}$ are combined with k-means clustering (15). An other way to perform clustering that have been explored consists in using d_0 or d_2 and hierarchical clustering (25).

The second category of nonparametric approaches proposes new heuristics to cluster functional data. For example, in (24), two dynamic programming algorithms perform simultaneously clustering and piecewise estimation of the cluster centers. In (53), a new procedure is developed to identify simultaneously optimal clusters of functions and optimal subspaces for clustering.

Model-based approaches

The first model-based clustering algorithm has been proposed in (30), under the name *fclust*. The authors consider that the expansion coefficients of the curves into a spline basis of functions are distributed according to a mixture Gaussian distributions with means μ_k, specific to each cluster and common variance Σ:

$$\alpha_i \sim \mathcal{N}(\mu_k, \Sigma).$$

Contrary to the two-stage approaches, in which the basis expansion coefficients are considered fixed, they are considered as random, what allows to proceed efficiently with sparsely sampled curves. Parsimony assumptions on the cluster means μ_k allow to define parsimonious clustering models and low-dimensional graphical representation of the curves.

The use of spline basis is convenient when the curves are regular, but are not appropriate for peak-like data as encountered in mass spectrometry for instance. For this reason, (22) recently proposes a Gaussian model on a wavelet decomposition of the curves, which allows to deal with a wider range of functional shapes than splines.

An interesting approach has also been considered in (47), by assuming that the curves arise from a mixture of regressions on a basis of polynomial functions, with possible changes in regime at each instant of time.

C.2.3 Model selection

Choosing the number of clusters

If classical model selection tools, as BIC, AIC or ICL are frequently used in the context of model-based clustering to select the number of clusters, more specific criteria have also been introduced.

First of all, Bayesian model for functional data clustering defines a framework in which the number of clusters can be directly estimated. For instance, (23) considered a uniform prior over the range $\{1, ..., n\}$ for the number of clusters, which is then estimated when maximizing the posterior distribution.

More empirical criteria have also been used for functional data clustering. In the two-stage clustering method presented in(33), the clustering is repeated several times for each number of clusters and that leading to the highest stability of the partition is retained. Even more empirical and very sensitive, (25) retain the number of clusters leading to a partition having the best physical interpretation.

In (30), an original model selection criterion is considered. This criterion is defined as the averaged Mahalanobis distance between the basis expansion coefficients and their closest cluster center.

Choosing the approximation basis

Almost all clustering algorithms for functional data needs the approximation of the curves into a finite dimensional basis of functions. Therefore, there is a need to choose an appropriate basis and thus, the number of basis functions. In (42), the authors advise to choose the basis according to the nature of the functional data: for instance, Fourier basis can be suitable for periodic data, whereas spline basis is the most common choice for non-periodic functional data. The other solution is to use less subjective criteria such as penalized likelihood criteria BIC, AIC or ICL.

C.3 Software

Whereas there exist several software solutions for finite dimensional data clustering, the software devoted to functional data clustering is less developed.

Under the **R** software environment, two-stage methods can be performed using for instance the functions *kmeans* or *hclust* of the *stats* package, combined with the distances available from the *fda* or *fda.usc* packages. Alternatively, several recent model-based clustering algorithms have been implemented by their authors and are available under different forms:

- **R** functions for *funHDDC* (8) and *funclust* (28) are available from request from their authors. An **R** package is available since 2013 on the CRAN website,

- an **R** function for *fclust* (30) is available directly from James's webpage,

- the package *curvclust* for **R** is probably the most finalized tool for curves clustering in **R** and implements the wavelets-based methods (22).

A MATLAB toolbox, *Curve Clustering Toolbox* (19), implements a family of two-stage clustering algorithms combining mixture of Gaussian models with spline or polynomial basis approximation.