Analyse de données bio-médicales Recherche reproductible et modélisation mathématique

Li-Thiao-Té Sébastien

LAGA UMR 7539, Université Paris 13

Habilitation à diriger des recherches
soutenue le 23 novembre 2021

Plan

Parcours scientifique

- 2005-2009 : Doctorat en Mathématiques Appliquées, ENS Cachan et Institut Pasteur.
  Caractérisation et Détection de Signaux Protéiques en Spectrométrie de Masse
- 2009-2010 : Post-doctorat dans l'Équipe Statistique et Génome, UMR 518 AgroParisTech/INRA
  Estimation du nombre d'espèces en métagénomique
- 2010- : Maître de conférences en mathématiques appliquées, LAGA UMR 7539, Université Paris 13

Parcours scientifique

2010- : Maître de conférences en mathématiques appliquées, LAGA UMR 7539, Université Paris 13

- 2014-2018 : traitement d'images biomédicales avec F Dibos, M Luong, JM Rocchisani, DV Tran (doctorant)
- 2014-2018 : programme interdisciplinaire Imageries du Vivant (Pres SPC, Universités P5, P7 et P13)
- 2014- traitement du signal en spectroscopie RPE avec Y Frapart, DN Tran (doctorant)
- 2018- traitement d'images et modèles EDP pour les maladies inflammatoires intestinales avec H Zaag, J Chaussard, X Treton, E Ogier-Denis, S Al Ali (doctorante)

Problématiques rencontrées

- travail collaboratif et dialogue interdisciplinaire
- pour les applications, difficulté de s'approprier les méthodes et logiciels mathématiques
- pour les mathématiques, difficulté de valider les hypothèses de modélisation et les résultats de l'analyse de données

Plan

Qu'est-ce que la reproductibilité ?

Obstacles à la reproductibilité numérique

- le générateur de nombres aléatoires
- l'environnement logiciel
- l'opérateur

Contrôler l'environnement logiciel

- Environnement = versions des librairies, logiciels, compilateurs, etc.
- variable et évolutif
- il faut le contrôler :
  - décrire : approches de type provenance, guix
  - fournir : machines virtuelles, container, packages snap/flatpak

Contrôler l'opérateur

L'opérateur doit :

- écrire ce qu'il fait sans erreur/oubli
  $=>$ remplacer l'opérateur par un "robot" programmable

- décrire ce qu'il fait pour qu'on le comprenne
  $=>$ documentation du programme avant le code source

Logiciel Lepton : format et documentation

```
Documentation
<<chunk_name options>>=
source code
@
```

Documentation

Code chunk 1 : «chunk_name»

```
source code
```

Logiciel Lepton : opérations

quatre opérations de base :

- `-write` permet d'écrire sur le disque
- `-exec interpreter` permet d'exécuter le code
- `-chunk format -output format` contrôle le format de sortie
- `<<chunk_ref>>` permet de réutiliser un autre bloc

Logiciel Lepton : python

Fonction $x \mapsto x^2$ en Python :

Code chunk 2 : «python»

```
[i**2 for i in range(0,3)]
```

Interpret with python3

```
[0, 1, 4]
```

Logiciel Lepton : Ocaml

Fonction $x \mapsto x^2$ en Ocaml :

Code chunk 3 : «ocaml»

```
let f = fun x -> x*x in
List.map f [0;1;2];;
```

Interpret with ocaml

```
- : int list = [0; 1; 4]
```

Logiciel Lepton : C/shell

Code chunk 4 : «main.c»

```c
#include <stdio.h>
int main() {
  printf("Hello world.\n");
}
```

Code chunk 5 : «shell»

```
gcc main.c
./a.out
```

Interpret with shell

```
Hello world.
```

Logiciel Lepton : état actuel

Code source libre publié sur Github :
https ://github.com/slithiaote/lepton

Publication associée :
Lepton : An automaton for Literate Executable Papers, 2019, Journal of Open Source Software, https ://doi.org/10.21105

Mais : Écrire un document compréhensible est difficile. Il faut du travail et de la méthode.

Next steps

Architecture d'un projet de recherche

- Assets
- Source code
- Scripts
- Artefacts
- Reports

Mise en place progressive dans mes projets de recherche.

Next steps

Propriétés supplémentaires :

- modularité : découper les projets en petits morceaux
- utilisation explicite des différents éléments
- incrémentalité : garder en mémoire les calculs
- intégration continue : tests exhaustifs
- exécution distante : garantie d'universalité

Plan

Analyse de données

Pour apporter une "preuve mathématique", il faut savoir :

- Que sont les données ? que sont les modèles ?
- Que sont les énoncés scientifiques ?
- Où est contenue l'information ?

À cause des problèmes de reproductibilité (bruit), on ne peut pas utiliser simplement la théorie de la démonstration.

Analyse de données : définitions

- $\mathcal{E}$ = ensemble des données
- L'information est portée par l'observation de certains éléments de $\mathcal{E}$, par opposition aux autres éléments.

- $\mathcal{M}$ = ensemble des modèles
- L'information est portée par les modèles compatibles, par opposition aux autres modèles possibles.

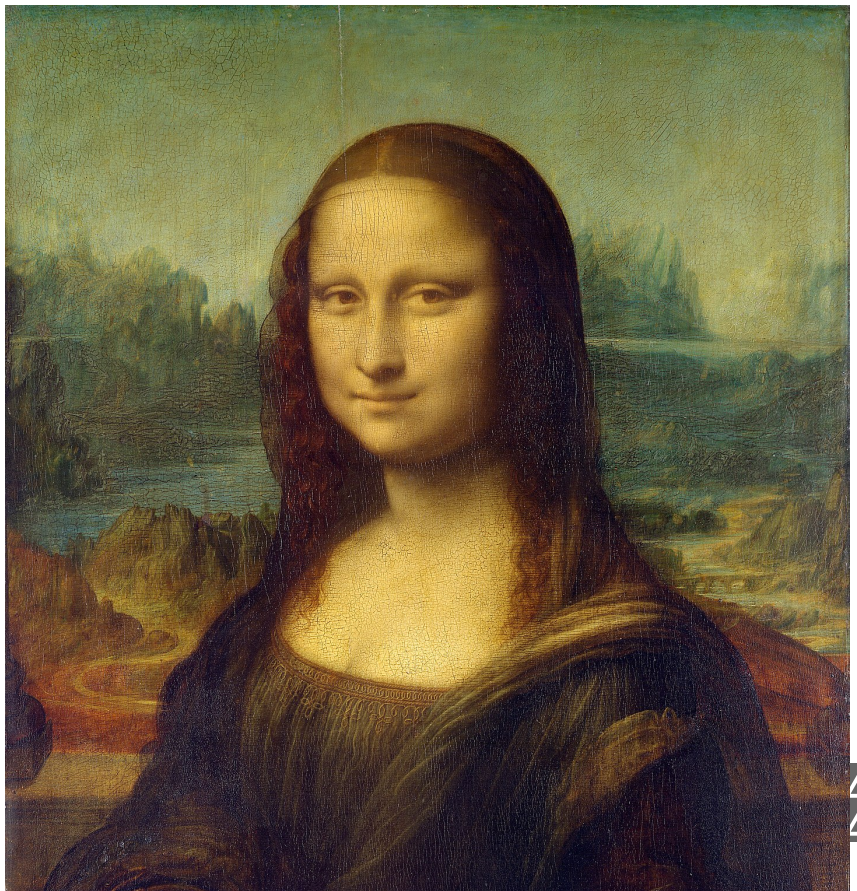
$\mathcal{M}$ et $\mathcal{E}$ sont reliés par : un modèle d'observation, une distance, une similarité, etc.

Analyse de données : reconnaissance faciale

- $\mathcal{M} =$ ensemble de noms
- noms compatibles avec l'image, par opposition aux autres noms possibles.

MONA $\mapsto$

- $\mathcal{E} =$ ensemble des images
- observation d'une certaine image, par opposition aux autres images possibles.



Problème direct et problème inverse

- Problème direct : le vrai modèle $m \in \mathcal{M}$ est connu.
  Par exemple générer des observations sous m.

  MONA $\mapsto$ 

- Problème inverse : les observations $e \in \mathcal{E}$ sont connues.
  On souhaite trouver le(s) modèle(s) correspondant(s), c'est-à-dire l'image réciproque de e dans $\mathcal{M}$ par l'opération du modèle d'observation.

  MONA, LISA, ADELE $\hookleftarrow$ 

Attention : dans beaucoup d'applications, l'espace $\mathcal{M}$, ainsi que le modèle d'observation sont inconnus et choisis arbitrairement.

Hypothèse de modélisation

C'est le choix de $\mathcal{E}$ ou de $\mathcal{M}$, ou de la relation entre $\mathcal{E}$ et $\mathcal{M}$.
$=>$ choix d'un sous-ensemble d'objets partageant des propriétés communes.

Cas particulier des propriétés imposées par l'observateur :

- elles sont justifiables
- invariance par groupes de transformation
  e.g. dans le cas des images : isométries, groupe projectif
- schéma d'expérience, échantillonnage

Hypothèse de modélisation

Certains choix conduisent à des problèmes triviaux.

Proposition

- on observe une loi empirique $\mathrm{d}\mathbb{P}_Y \in \mathcal{E}$,
- $\mathcal{M}$ est l'ensemble des variables aléatoires à valeurs dans $\mathcal{Y}$
- similarité définie par l'espérance $\mathbb{E}[d(M, Y)]$

Alors

$$\mathbb{E}[d(M, Y)] \geq \inf_{m \in \mathcal{Y}} \int d(m, y)\, \mathrm{d}\mathbb{P}_Y$$

En particulier, la masse de Dirac en m_0 est le modèle d'espérance minimale donc de similarité maximale.

Démonstration : Conséquence de l'indépendance entre modèles et observations

Énoncé scientifique

C'est un sous-ensemble $\mathcal{S} \subset \mathcal{E}$ ou $\mathcal{S} \subset \mathcal{M}$:

- point $\rightarrow$ estimation paramétrique,
- un ensemble / fonction indicatrice $\rightarrow$ test statistique,
- une partition de $\mathcal{E}$ ou de $\mathcal{M} \rightarrow$ classification.

Un énoncé inconnu peut être appris par un algorithme de machine learning.

Un énoncé connu peut être démontré par accumulation

- d'observations $e \in \mathcal{S}$
- ou de modèles $m \in \mathcal{S}'$

Preuves mathématiques

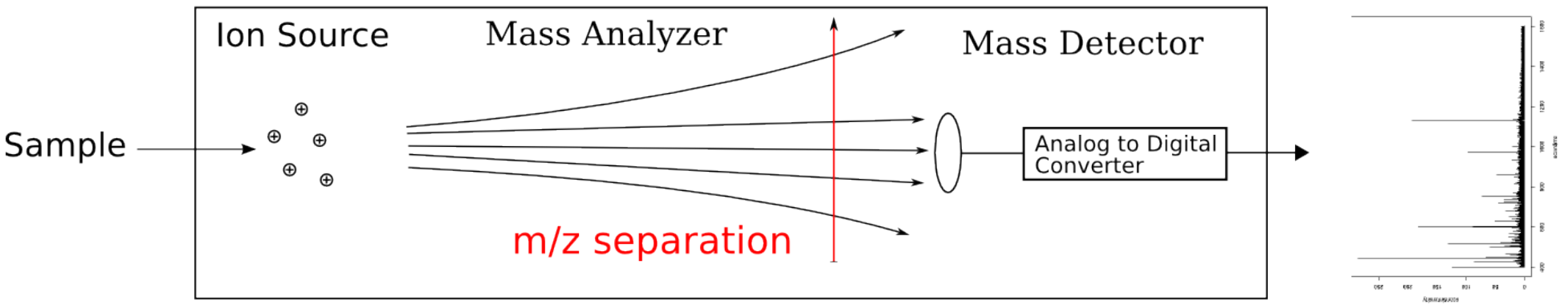
Le processus d'accumulation aboutit à un ensemble de modèles compatibles ou à un paysage dans l'ensemble des modèles.

On montre que

- l'ensemble des modèles compatibles est calculable,
- il fournit une analyse complète,
- on peut en déduire des énoncés intéressants.

Plan

Exemple inspiré d'un spectromètre de masse



Exemple inspiré d'un spectromètre de masse

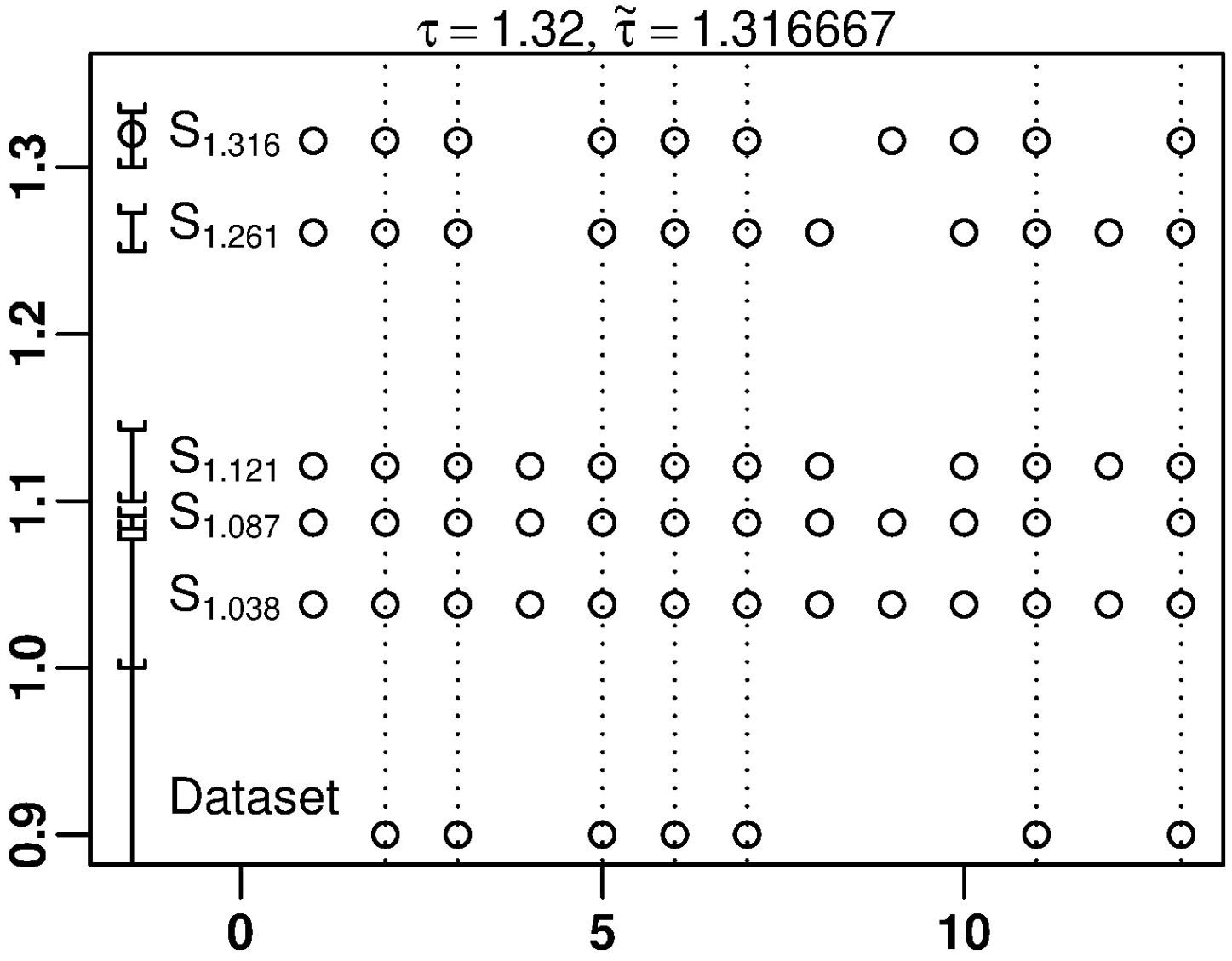
On considère le modèle

$$Y = \lfloor \tau X \rfloor$$

où le signal observé Y est obtenu à partir d'un nombre d'ions X après amplification d'un facteur de gain τ et troncature.

- Espace des données $\mathcal{E}$ = sous-ensembles de $\mathbb{N}$
- Espace des modèles $\mathcal{M} = \{\tau \in \mathbb{R}\}$
  <=> sous-ensemble généré $\mathcal{S}_\tau = \lfloor \tau \mathbb{N} \rfloor \subset \mathbb{N}$
- Un modèle τ est compatible si l'observation $Y \subset \mathcal{S}_\tau$
  i.e. aucune contradiction dans les observations.

Modèles compatibles avec les données



Résultats

L'ensemble des modèles compatibles est calculable explicitement :

- c'est une union d'intervalles
- majorée par $\frac{\max Y+1}{n}$ où n est le nombre d'entiers distincts observés
- la largeur des intervalles est minorée par $\frac{1}{(\max Y)^2}$

On peut retrouver tous les intervalles par échantillonnage.

Résultats

L'ensemble des modèles compatibles fournit un estimateur précis pour τ :

- tous les modèles sont acceptables
- les grandes valeurs de τ sont cohérentes avec un faible nombre de données manquantes
- on prend $\hat{\tau}$ la plus grande valeur compatible
- on échantillonne en descendant à partir du majorant
- mieux : on prend $\hat{\tau} = \frac{a+b}{2}$ où $[a, b[$ est le dernier intervalle de l'ensemble des valeurs compatibles

Énoncés scientifiques

Un énoncé est une partie de $\mathcal{E}$ ou de $\mathcal{M}$:

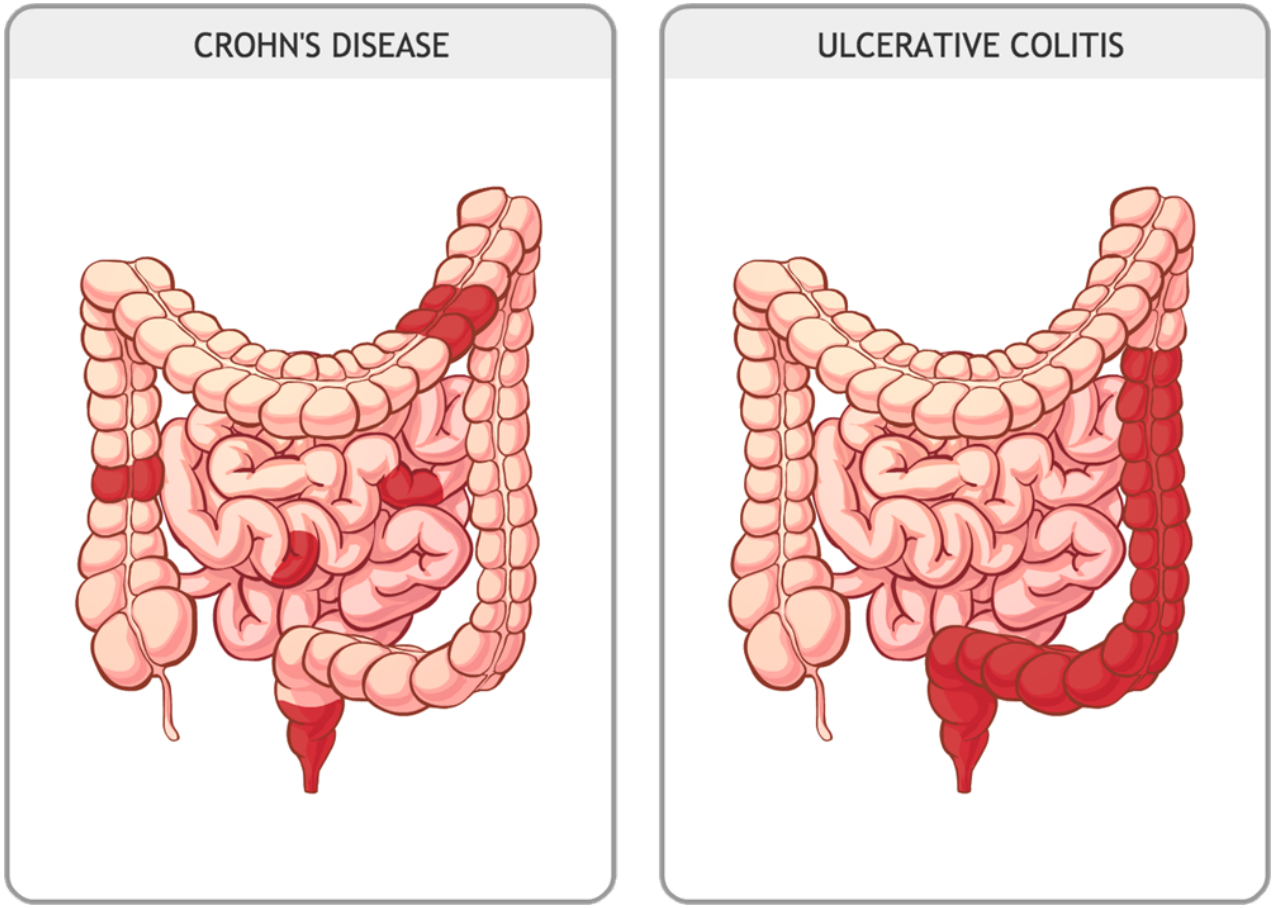
- "est-ce que l'on observe uniquement des nombres entiers ?" i.e. τ multiple de 2
- "est-ce que τ est égal à 3.14 ?"
- "est-ce que τ appartient à un l'intervalle $[1, 2]$?"

On démontre par accumulation

- on a plusieurs observations et toutes vérifient l'énoncé
- tous les modèles compatibles vérifient l'énoncé

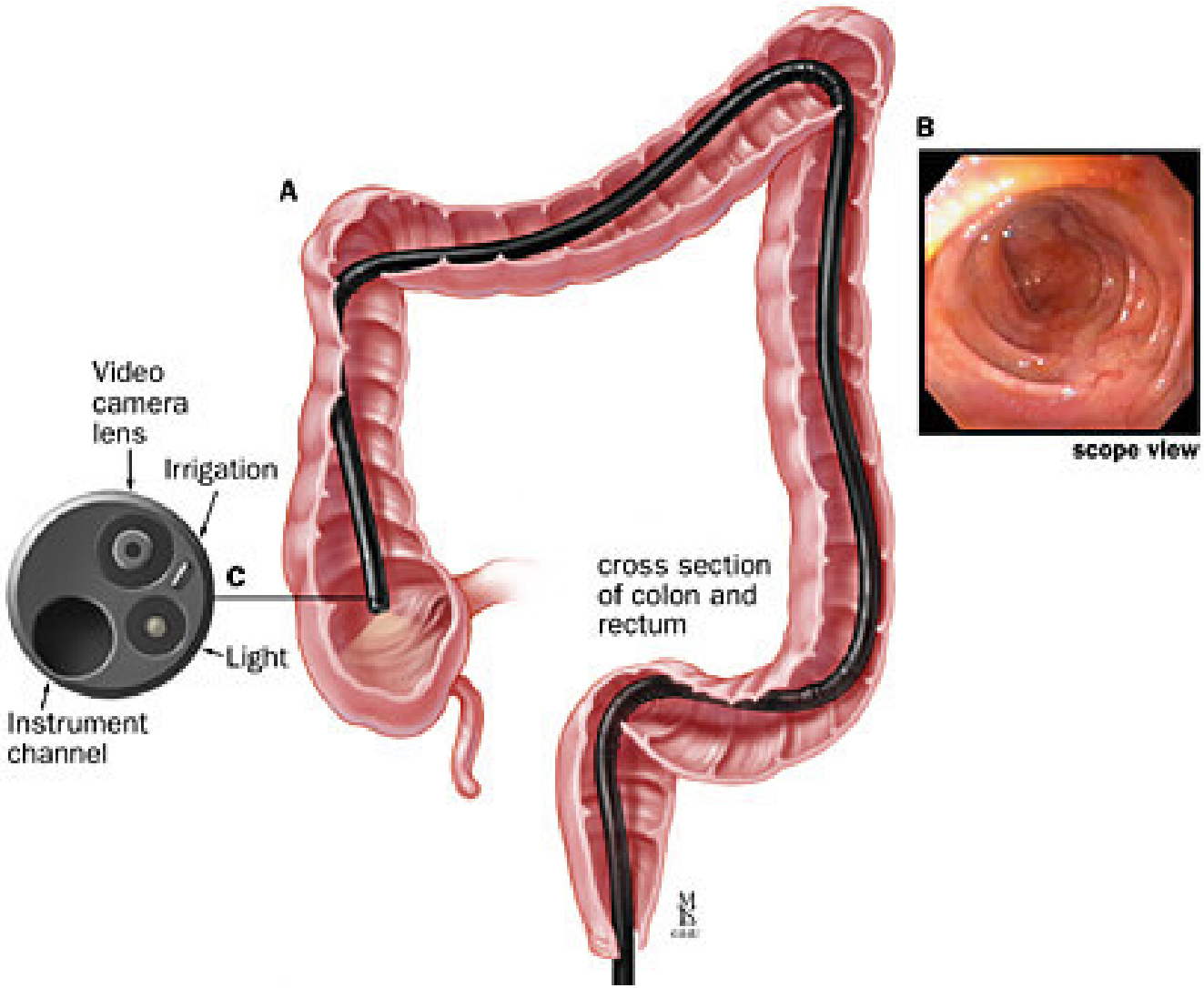
Plan

Maladies inflammatoires chroniques de l'intestin

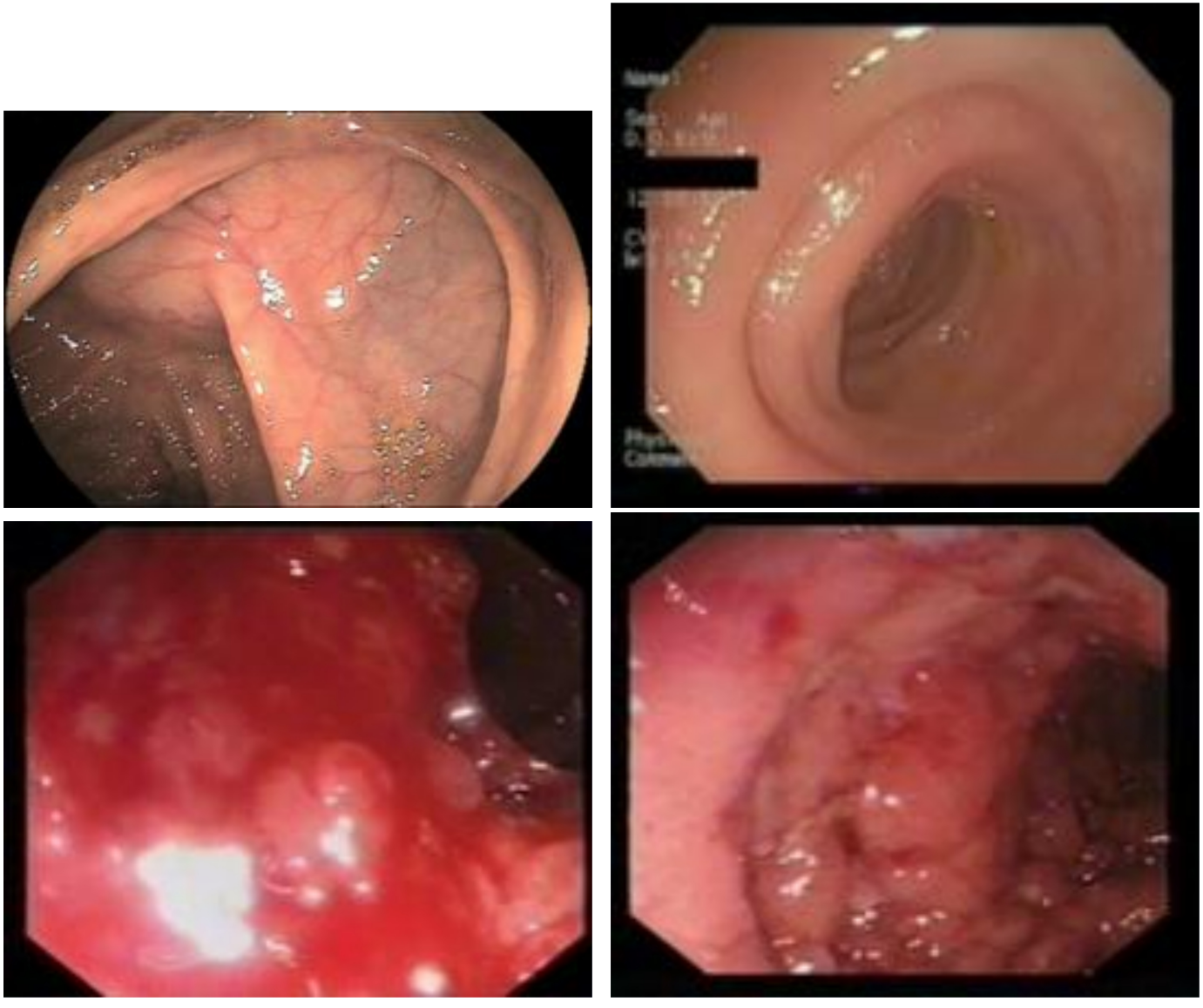


- Collaboration avec H Zaag, J Chaussard, X Treton, E Ogier-Denis
- Encadrement de la thèse de Safaa Al Ali
- Analyse d'image
- Modélisation par EDP

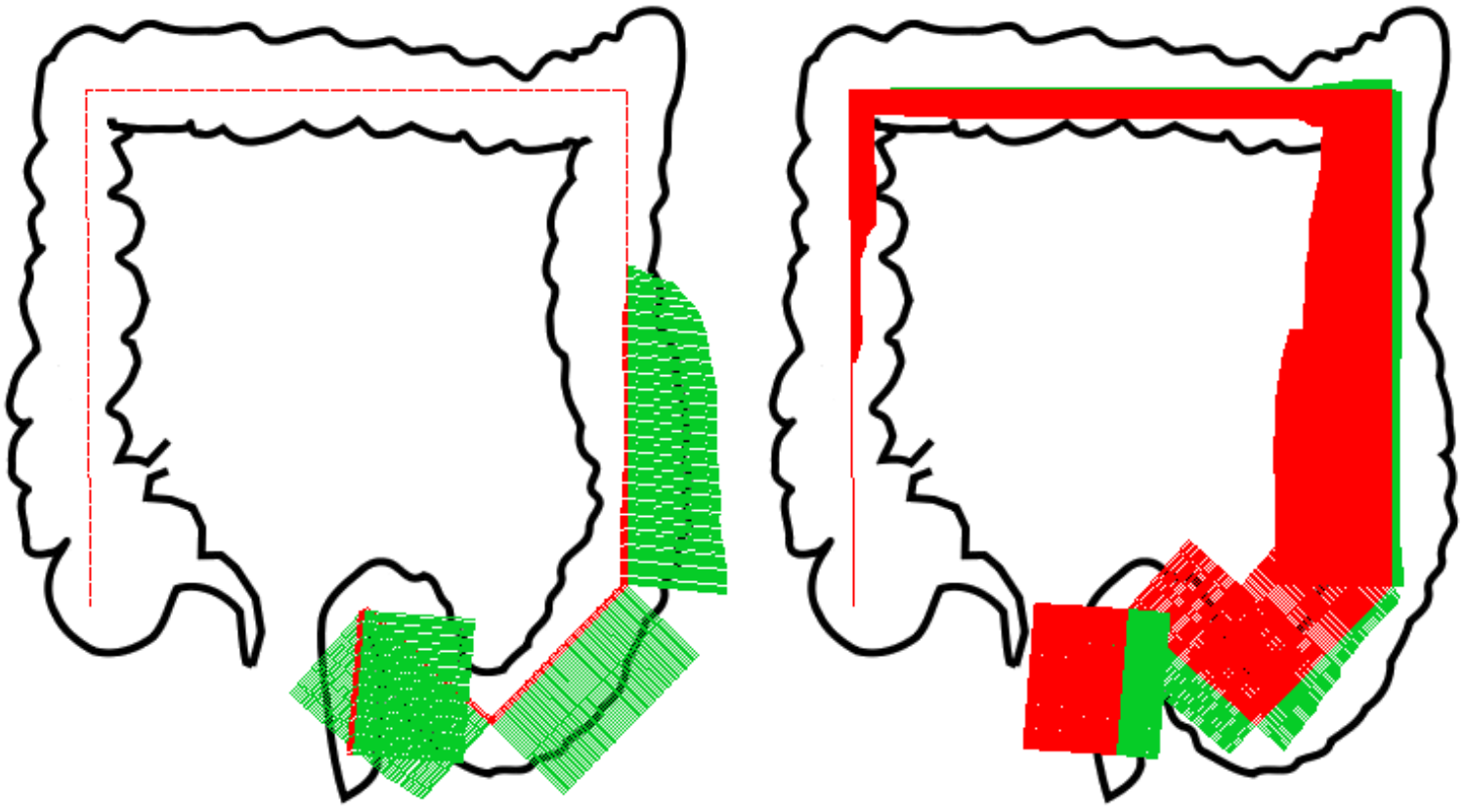
Vidéos de coloscopie



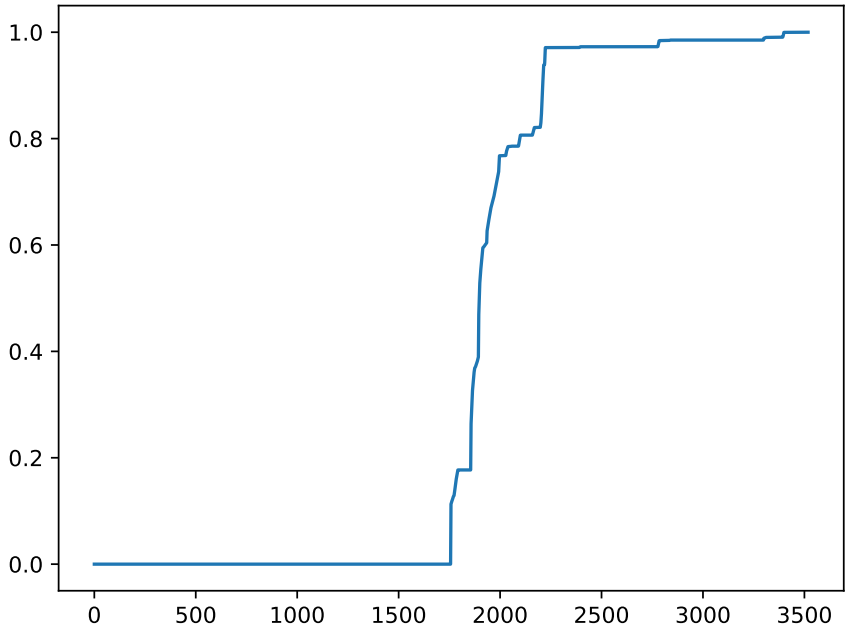
Lésions de la paroi intestinale



Profil des lésions



Distribution spatiale des lésions



- une observation est un profil $p(s) : [0, 1] \to \mathbb{R}$ (un patient)
- un modèle est une fonction $u(s) : [0, 1] \to \mathbb{R}$
- hypothèse 1 : distance L_2 entre $\mathcal{E}$ et $\mathcal{M}$
- hypothèse 2 : $\mathcal{M}$ est restreint à l'ensemble des solutions d'une équation de réaction-diffusion (FKPP)

Équation de FKPP

$$\frac{\partial}{\partial t}u - D\frac{\partial^2}{\partial s^2}u = u(1-u), \quad t \geq 0,\ s \in \mathbb{R}$$

L'espace $\mathcal{M}$ est paramétré par les couples (D, u_0) où D est le paramètre de diffusion et u_0 la condition initiale.

Résultats

L'espace des modèles est réduit à une dimension.

- les patients sont observés en régime asymptotique
- Pour $D = 1$, le profil de la solution de FKPP converge vers un front d'onde W_c de vitesse 2, indépendant de u_0.
- Les autres fronts s'obtiennent à partir de W_c par dilatation

Conclusion : $\mathcal{M} = \left\{ \sqrt{D}.W_c \quad \text{pour} \quad D \in \mathbb{R}^+ \right\}$

Résulats

On obtient l'ensemble des modèles compatibles par recalage.

- recalage à translation et dilatation près
- on échantillonne les modèles, i.e. les valeurs de D
- le meilleur modèle $D^* = \arg\min L_2(D, \text{patient})$
- les modèles compatibles
  $\{D \quad \text{t.q.} \quad L_2(D, \text{patient}) < 1.05\, L_2(D^*, \text{patient})\}$

Résultats

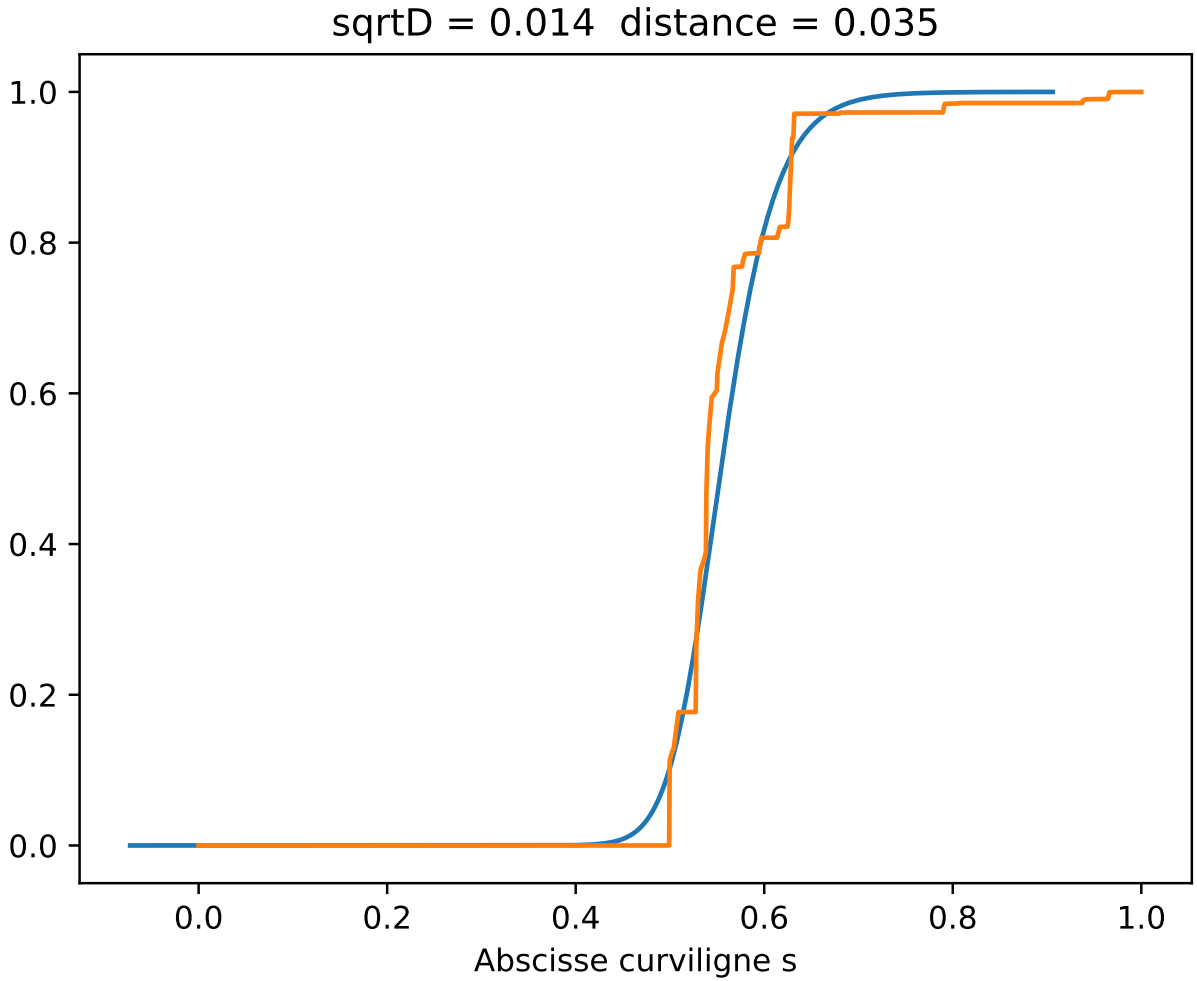


Figure – Recalage du front d'onde $\sqrt{D}.W_c$ sur le patient 23 avec un coefficient de diffusion $D = 0.014^2$.

Résultats

On peut calculer la vitesse de propagation du front $\sqrt{D}.W_c$

- le front se propage à une vitesse $v = 2\sqrt{D}$
- on obtient l'ensemble des vitesses compatibles
- ou la durée d'invasion

Code chunk 6 : «fkpp_recalage (part 4)»

```
D = np.arange(0.01,0.02,0.0005) ** 2
dists = [fkpp_interp(p,d) for d in D]
alpha, min_dist = 0.05, np.amin(dists)
scompat = 2 * np.sqrt(np.extract(dists<(1+alpha)*min_dist,D))
print('Vitesses compatibles', scompat)
Xp = np.linspace(0,1,num=len(p))
print('Temps d\'invasion',np.round((1-np.trapz(p,dx=Xp[1]-Xp[0]))/scompa
```

Interpret with python3

```
Vitesses compatibles [0.025 0.026 0.027 0.028 0.029 0.03  0.031]
Temps d'invasion [22.39 21.53 20.73 19.99 19.3  18.66 18.06]
```

Perspectives

- une observation est un profil $[0,1] \to \mathbb{R}$ (pour un patient)
  $=>$ plusieurs patients
- un modèle est une fonction $[0,1] \to \mathbb{R}$
  $=>$ autres types de modèles
- hypothèse 1 : distance L_2 entre $\mathcal{E}$ et $\mathcal{M}$
  $=>$ autres distances
- hypothèse 2 : $\mathcal{M}$ est restreint à l'ensemble des solutions d'une équation de réaction-diffusion (FKPP)
  $=>$ autres sous-espaces de fonctions

Conclusion

Un travail sur les outils pour l'analyse de données

- Logiciel Lepton
- Mise en oeuvre de la recherche reproductible
  - documents exécutables, faciles à réutiliser
  - documents complets, transparents, faciles à comprendre
- Ce n'est pas uniquement un problème "logiciel"
  - documents de qualité supérieure
  - changements de pratique, compétence à acquérir
- Perspectives
  - recommandations sur l'architecture des projets
  - propriétés supplémentaires : modularité, incrémentalité, exécution distante, etc.
  - proposition d'un module de formation doctorale

Conclusion

Un travail sur la notion de modèle et d'énoncé scientifique :

- il faut modéliser
  - les données observées, sous-ensemble des données possibles $\mathcal{E}$,
  - les modèles compatibles, sous-ensemble de l'ensemble des modèles $\mathcal{M}$.
- l'analyse de données est un travail sur des ensembles
  - énoncé scientifique = sous-ensemble de $\mathcal{E}$ ou de $\mathcal{M}$,
  - élucider l'ensemble des modèles compatibles et leurs propriétés

Conclusion

Un travail sur la notion de modèle et d'énoncé scientifique :

- C'est faisable, par exemple
  - modèles compatibles
  - propriétés de l'ensemble des images (thèse de DV Tran)
  - propriétés des modèles de spectres RPE (thèse de DN Tran)
  - modélisation de l'aspect des lésions de la RCH et de leur répartition spatiale (thèse de S Al Ali)
- Perspectives
  - zoologie des modèles, en fonction du type des données
  - mise en relation avec les autres thématiques : machine learning, statistiques, etc.
  - énumération des hypothèses pour la chimie