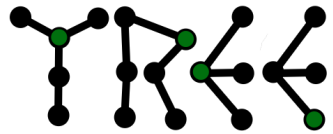
Approches algorithmiques pour la statistique :



Romain Azaïs

Soutenance d'habilitation à diriger des recherches

2 décembre 2022

Domaine de recherche : mathématiques appliquées $\supset$ statistique

Domaine de recherche : mathématiques appliquées $\supset$ statistique

Sujet : comprendre les enjeux statistiques d'un modèle stochastique

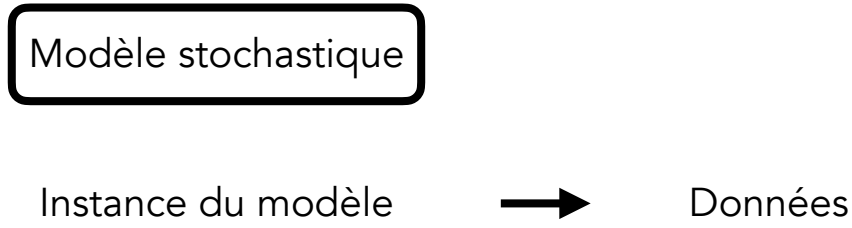
Domaine de recherche : mathématiques appliquées $\supset$ statistique

Sujet : comprendre les enjeux statistiques d'un modèle stochastique

Modèle stochastique

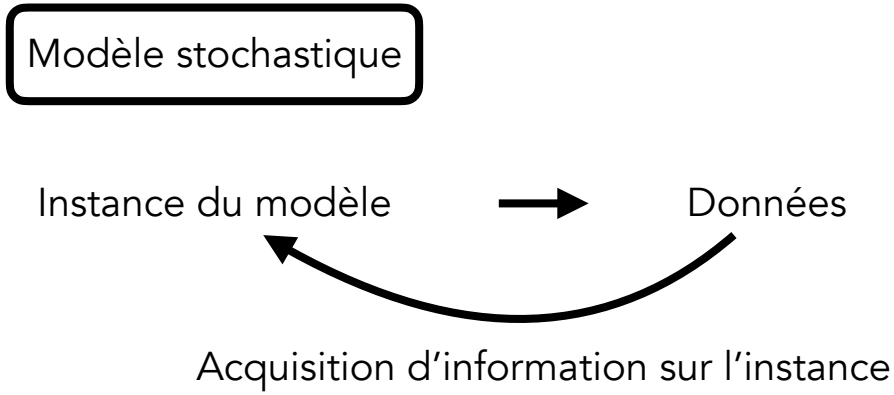
Domaine de recherche : mathématiques appliquées $\supset$ statistique

Sujet : comprendre les enjeux statistiques d'un modèle stochastique



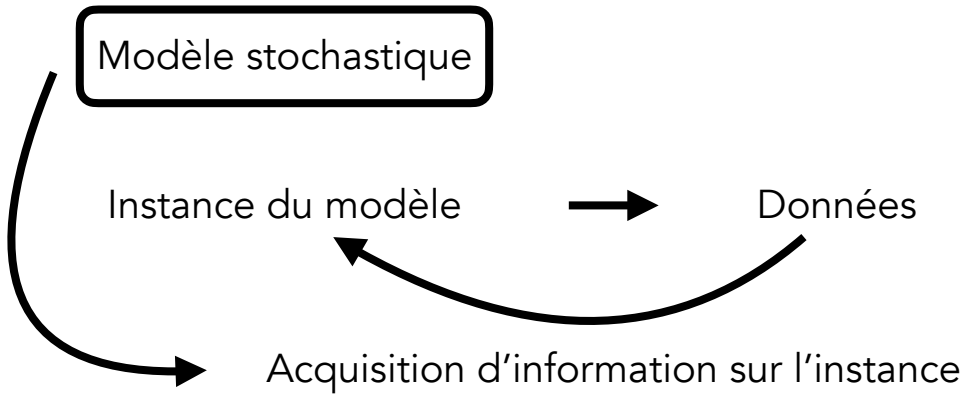
Domaine de recherche : mathématiques appliquées $\supset$ statistique

Sujet : comprendre les enjeux statistiques d'un modèle stochastique



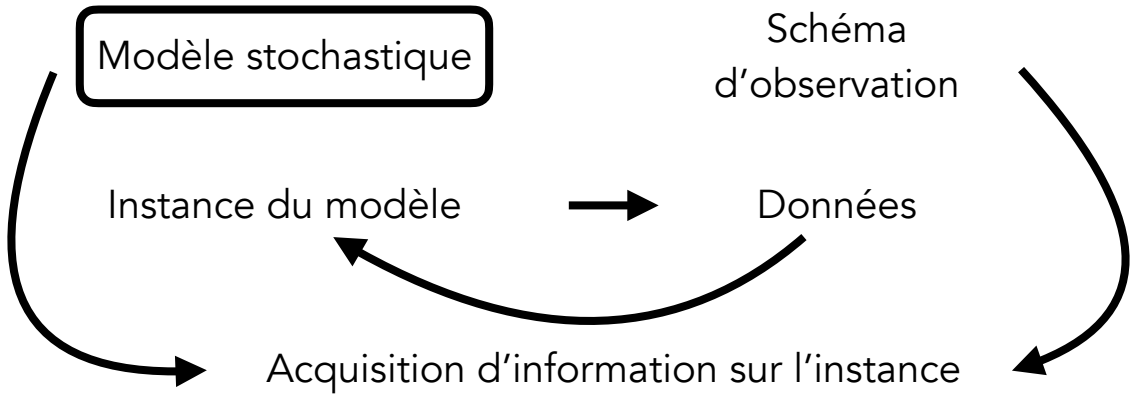
Domaine de recherche : mathématiques appliquées $\supset$ statistique

Sujet : comprendre les enjeux statistiques d'un modèle stochastique



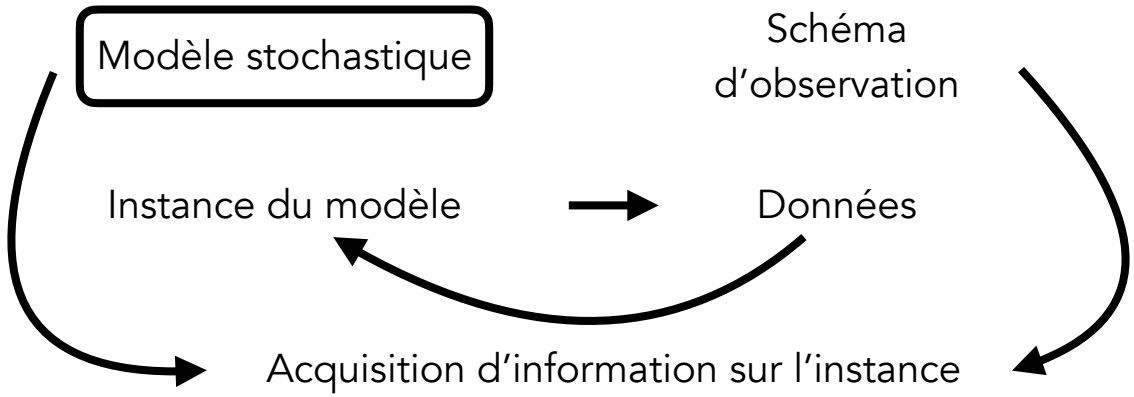
Domaine de recherche : mathématiques appliquées $\supset$ statistique

Sujet : comprendre les enjeux statistiques d'un modèle stochastique



Domaine de recherche : mathématiques appliquées $\supset$ statistique

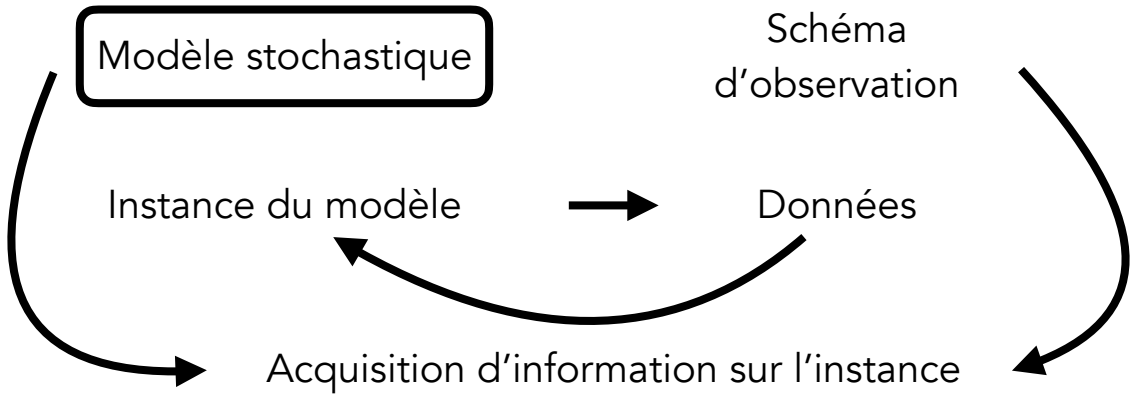
Sujet : comprendre les enjeux statistiques d'un modèle stochastique



Approche : construction d'algorithmes d'extraction de l'information adaptés au modèle dans sa généralité et théoriquement fondés

Domaine de recherche : mathématiques appliquées $\supset$ statistique

Sujet : comprendre les enjeux statistiques d'un modèle stochastique

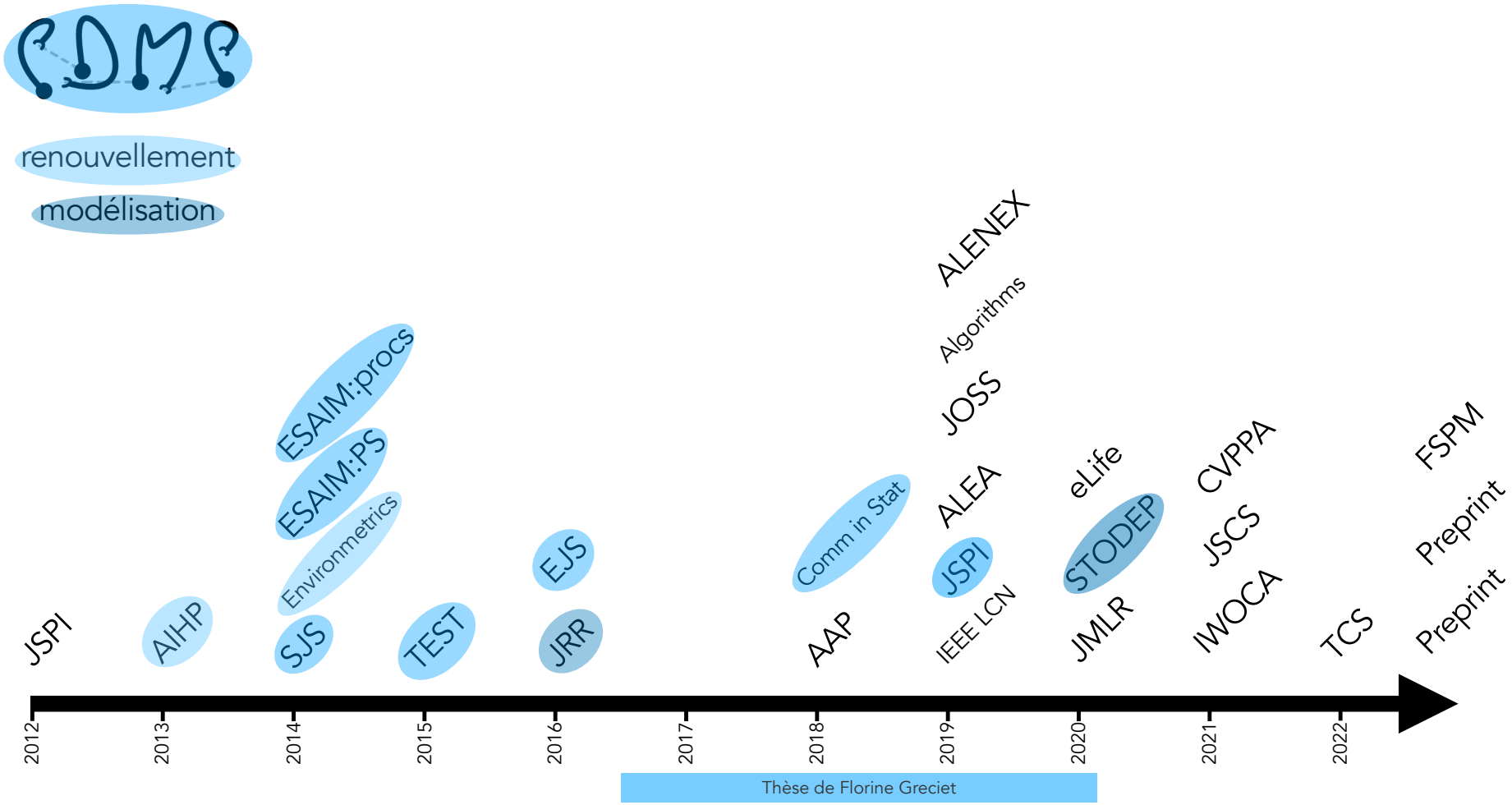


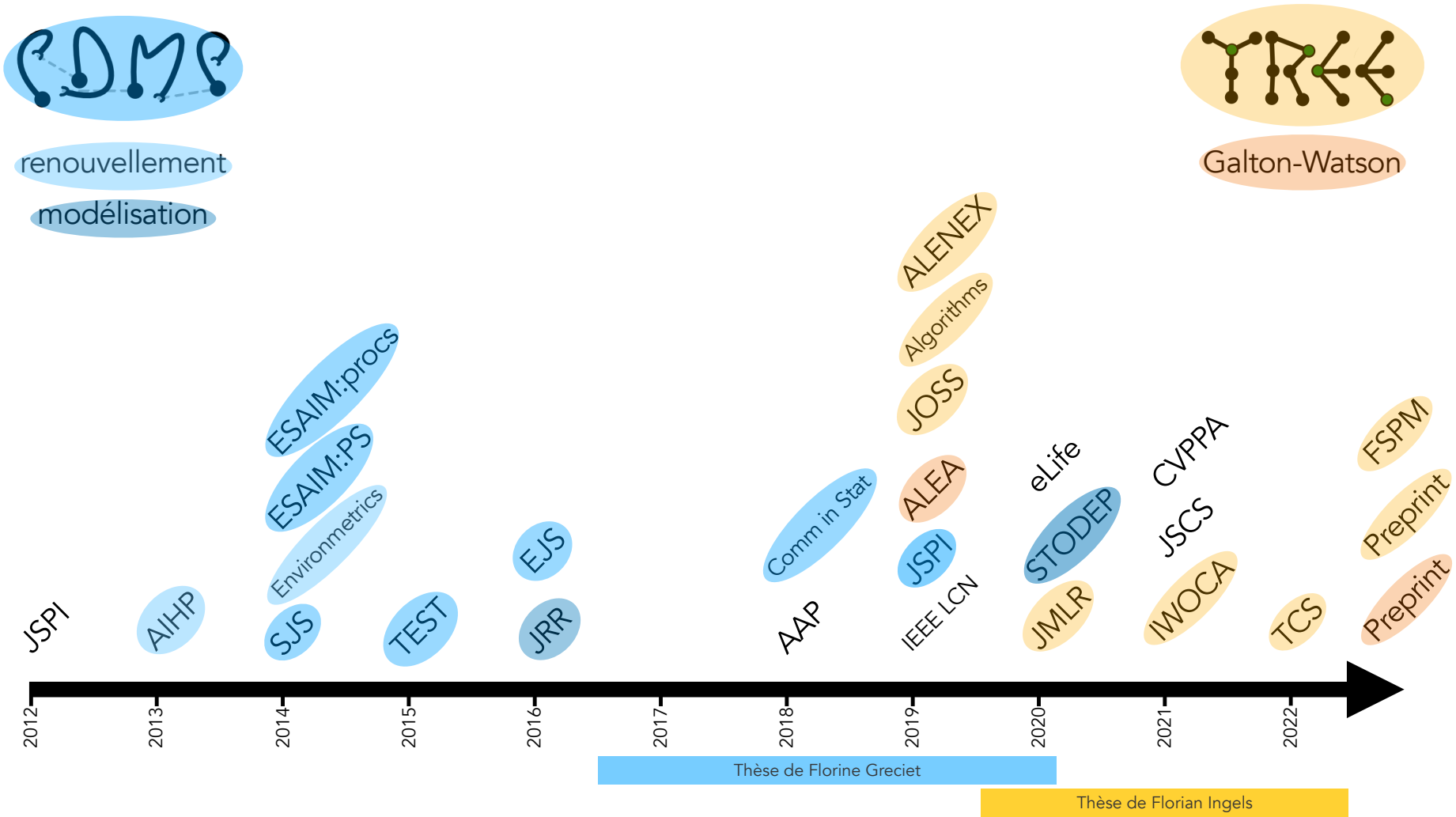
Approche : construction d'algorithmes d'extraction de l'information adaptés au modèle dans sa généralité et théoriquement fondés

Intérêt de l'implémentation :

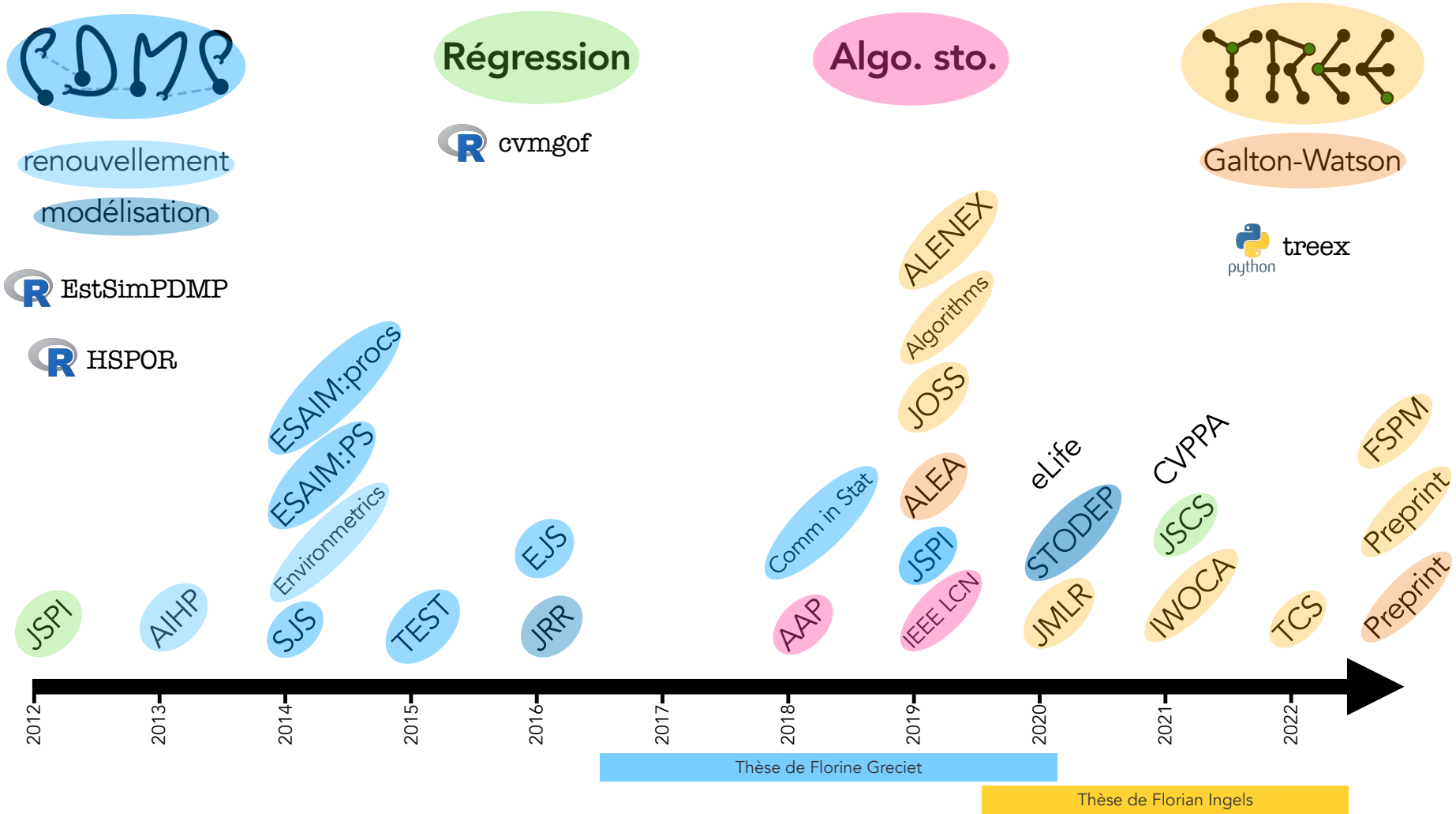
- donner des idées
- illustrer et/ou valider le comportement de l'algorithme
- aller vers l'applicabilité

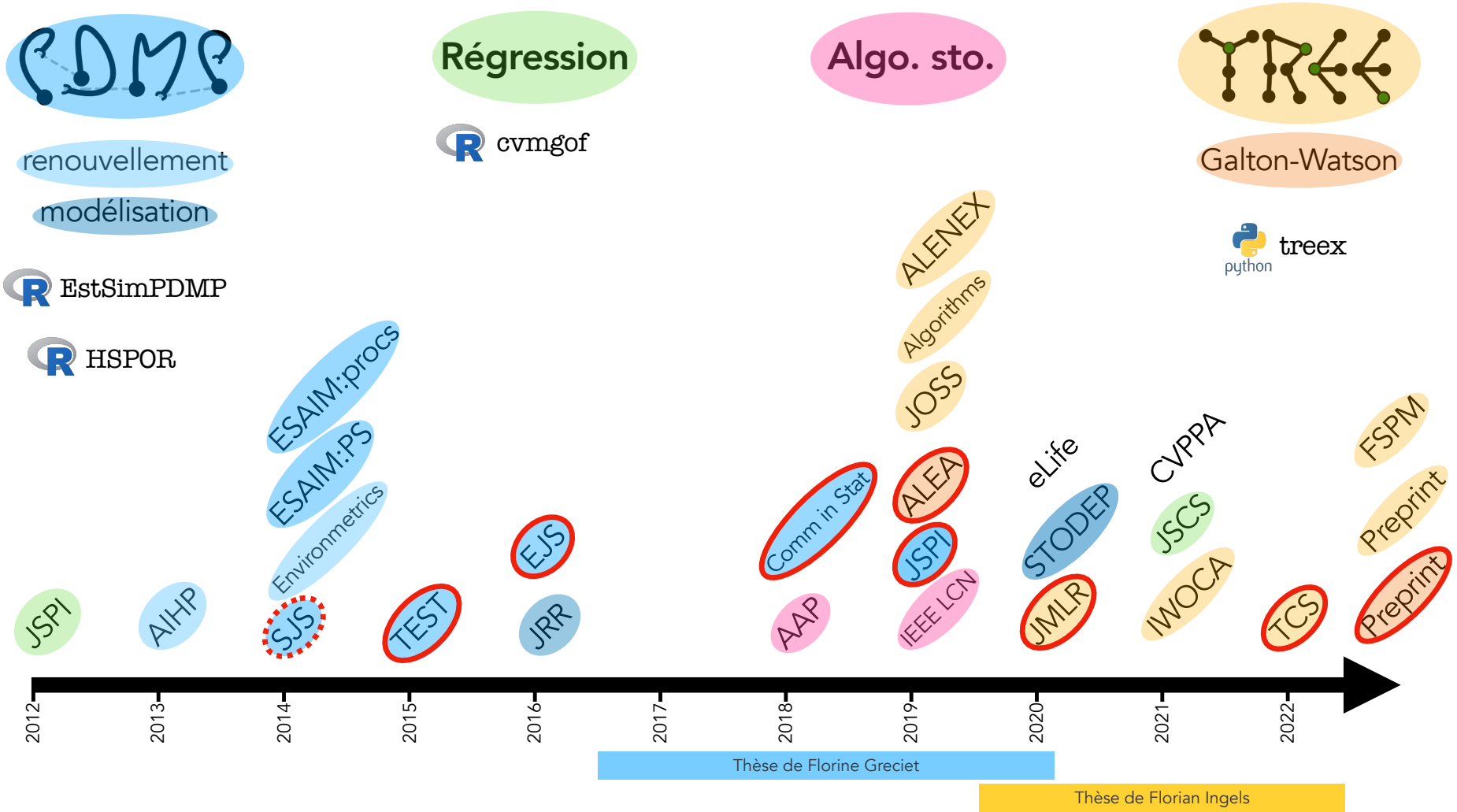
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
JSPI
AIHP
ESAIM:procs
ESAIM:PS
Environmetrics
SJS
TEST
EJS
JRR
Comm in Stat
AAP
ALENEX
Algorithms
JOSS
ALEA
JSPI
IEEE LCN
eLife
STODEP
JMLR
CVPPA
JSCS
IWOCA
TCS
FSPM
Preprint
Preprint


renouvellement
modélisation
JSPI
AIHP
ESAIM:procs
ESAIM:PS
Environmetrics
SJS
TEST
EJS
JRR
Comm in Stat
AAP
ALENEX
Algorithms
JOSS
ALEA
JSPI
IEEE LCN
eLife
STODEP
JMLR
CVPPA
JSCS
IWOCA
TCS
FSPM
Preprint
Preprint
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Thèse de Florine Greciet


renouvellement
modélisation
Galton-Watson
ALENEX
Algorithms
JOSS
ALEA
JSPI
IEEE LCN
eLife
STODEP
JMLR
CVPPA
JSCS
IWOCA
TCS
FSPM
Preprint
Preprint
Comm in Stat
AAP
ESAIM:procs
ESAIM:PS
Environmetrics
SJS
TEST
EJS
JRR
JSPI
AIHP
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Thèse de Florine Greciet
Thèse de Florian Ingels


renouvellement
modélisation
Régression
Algo. sto.
Galton-Watson
ALENEX
Algorithms
JOSS
ALEA
JSPI
IEEE LCN
eLife
STODEP
JMLR
CVPPA
JSCS
IWOCA
TCS
FSPM
Preprint
Preprint
Comm in Stat
AAP
EJS
JRR
ESAIM:procs
ESAIM:PS
Environmetrics
SJS
TEST
JSPI
AIHP
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Thèse de Florine Greciet
Thèse de Florian Ingels


renouvellement
modélisation
EstSimPDMP
HSPOR
Régression
cvmgof
Algo. sto.
Galton-Watson
treex
python
ESAIM:procs
ESAIM:PS
Environmetrics
SJS
JSPI
AIHP
TEST
EJS
JRR
Comm in Stat
AAP
ALENEX
Algorithms
JOSS
ALEA
JSPI
IEEE LCN
eLife
STODEP
JMLR
CVPPA
JSCS
IWOCA
TCS
FSPM
Preprint
Preprint
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Thèse de Florine Greciet
Thèse de Florian Ingels

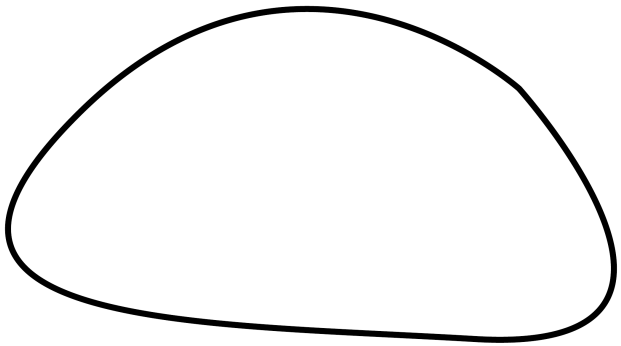

renouvellement
modélisation
EstSimPDMP
HSPOR
Régression
cvmgof
Algo. sto.
Galton-Watson
treex
python
ALENEX
Algorithms
JOSS
ALEA
JSPI
IEEE LCN
Comm in Stat
AAP
eLife
STODEP
JMLR
CVPPA
JSCS
IWOCA
TCS
FSPM
Preprint
Preprint
ESAIM:procs
ESAIM:PS
Environmetrics
SJS
JSPI
AIHP
TEST
EJS
JRR
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Thèse de Florine Greciet
Thèse de Florian Ingels

Processus markoviens déterministes par morceaux (PDMPs)

- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP

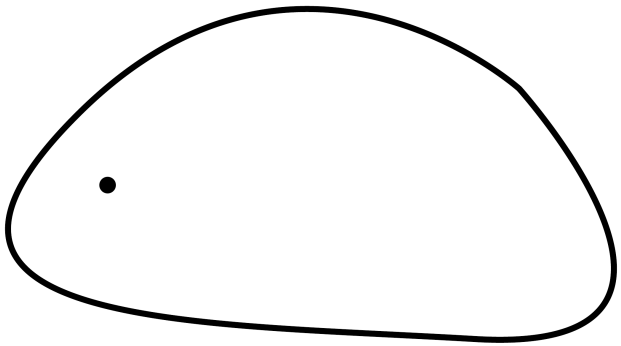
Processus markoviens déterministes par morceaux (PDMPs)

- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



Processus markoviens déterministes par morceaux (PDMPs)

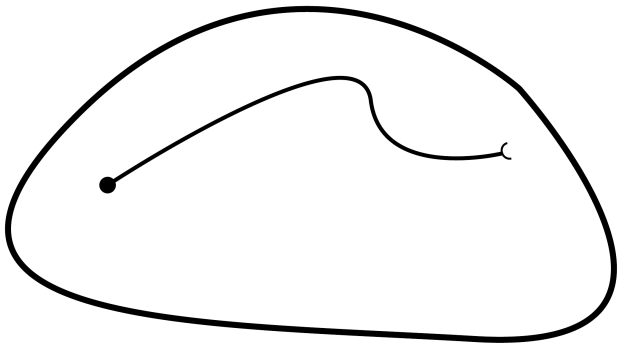
- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



- Condition initiale : $X_0 = x$

Processus markoviens déterministes par morceaux (PDMPs)

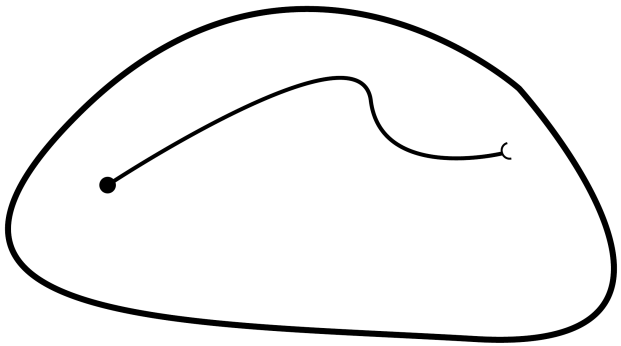
- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



- Condition initiale : $X_0 = x$
- Promenade déterministe :
  $\forall\, 0 \leq t < T_1,\ X_t = \Phi(t|x)$

Processus markoviens déterministes par morceaux (PDMPs)

- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP

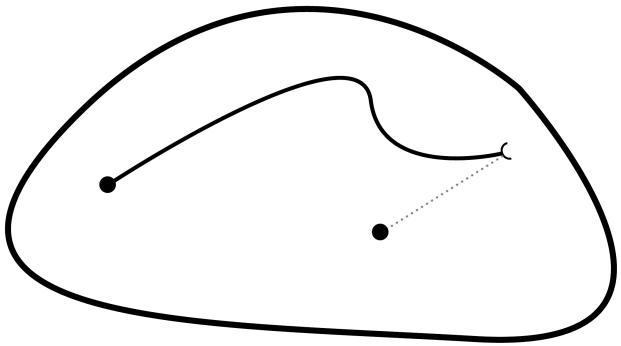


- Condition initiale : $X_0 = x$
- Promenade déterministe :
  $\forall\, 0 \leq t < T_1,\ X_t = \Phi(t|x)$
- Durée aléatoire :

$$\mathbb{P}(T_1 > t) = \exp\left(-\int_0^t \lambda\left(\Phi(s|x)\right) \mathrm{d}s\right) \mathbb{1}_{\{0 \leq t < t^+(x)\}}$$

Processus markoviens déterministes par morceaux (PDMPs)

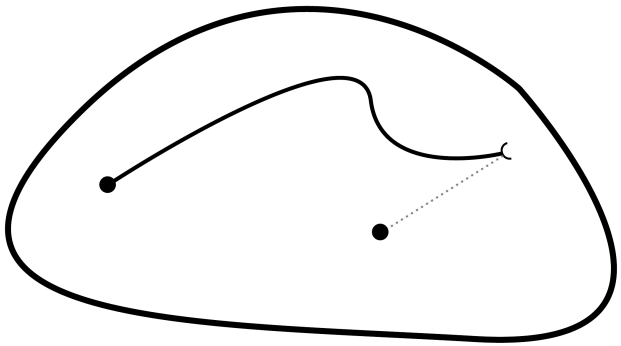
- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



- Condition initiale : $X_0 = x$
- Promenade déterministe :
  $\forall\, 0 \leq t < T_1,\ X_t = \Phi(t|x)$
- Durée aléatoire :
  $$\mathbb{P}(T_1 > t) = \exp\left(-\int_0^t \lambda\left(\Phi(s|x)\right) \mathrm{d}s\right) \mathbb{1}_{\{0 \leq t < t^+(x)\}}$$
- Perturbation aléatoire :
  $$\mathbb{E}\left[\varphi(X_{T_1}) \,|\, \Phi(T_1|x)\right] = \int \varphi(u) Q(\mathrm{d}u|\Phi(T_1|x))$$

Processus markoviens déterministes par morceaux (PDMPs)

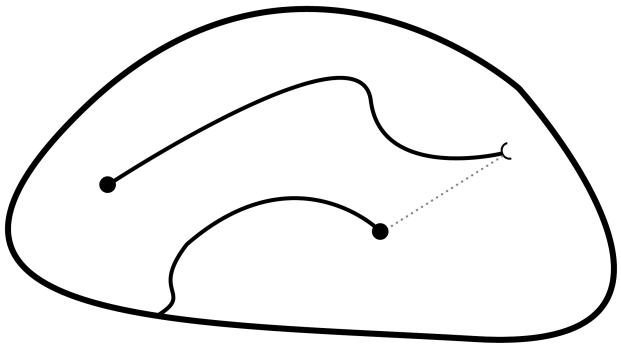
- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



- $X_{T_1} = x$

Processus markoviens déterministes par morceaux (PDMPs)

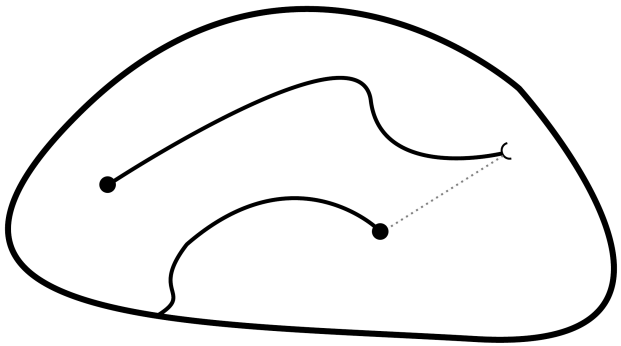
- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



- $X_{T_1} = x$
- Promenade déterministe :
  $\forall\, 0 \leq t < S_2,\ X_{T_1+t} = \Phi(t|x)$

Processus markoviens déterministes par morceaux (PDMPs)

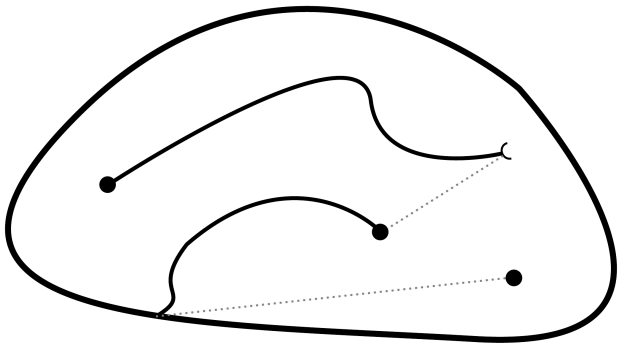
- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



- $X_{T_1} = x$
- Promenade déterministe :
  $\forall\, 0 \leq t < S_2,\ X_{T_1+t} = \Phi(t|x)$
- Durée aléatoire :
  $$\mathbb{P}(S_2 > t) = \exp\left(-\int_0^t \lambda\left(\Phi(s|x)\right) \mathrm{d}s\right) \mathbb{1}_{\{0 \leq t < t^+(x)\}}$$

Processus markoviens déterministes par morceaux (PDMPs)

- Introduits dans les années 80 par Davis
- The class is "wide enough to include as special cases virtually all the non-diffusion models of applied probability"
- MC, CT-MC, M/G/1 queue, G/G/1 queue $\subset$ PDMP



- $X_{T_1} = x$
- Promenade déterministe :
  $\forall\, 0 \leq t < S_2,\ X_{T_1+t} = \Phi(t|x)$
- Durée aléatoire :
  $$\mathbb{P}(S_2 > t) = \exp\left(-\int_0^t \lambda\left(\Phi(s|x)\right) \mathrm{d}s\right) \mathbb{1}_{\{0 \leq t < t^+(x)\}}$$
- Perturbation aléatoire :
  $$\mathbb{E}\left[\varphi(X_{T_2}) \,|\, \Phi(S_2|x)\right] = \int \varphi(u) Q(\mathrm{d}u|\Phi(S_2|x))$$

Statistique des PDMPs

- Reconstruction des trajectoires ($\simeq$ estimation de Φ)
- Estimation des caractéristiques probabilistes (λ et Q)
- Estimation de fonctionnelles

Statistique des PDMPs

- Reconstruction des trajectoires ($\simeq$ estimation de Φ)

  Méthodes heuristiques pour des problèmes appliqués
  JRR '16, thèse de Florine Greciet

- Estimation des caractéristiques probabilistes (λ et Q)

- Estimation de fonctionnelles

Statistique des PDMPs

- Reconstruction des trajectoires ($\simeq$ estimation de Φ)

  Méthodes heuristiques pour des problèmes appliqués
  JRR '16, thèse de Florine Greciet

- Estimation des caractéristiques probabilistes (λ et Q)

  Statistique non-paramétrique dans un cadre général
  ESAIM:PS '13, SJS '14 , EJS '16 , Comm in Stat '18

  thèse

- Estimation de fonctionnelles

Statistique des PDMPs

- Reconstruction des trajectoires ($\simeq$ estimation de Φ)

  Méthodes heuristiques pour des problèmes appliqués
  JRR '16, thèse de Florine Greciet

- Estimation des caractéristiques probabilistes (λ et Q)

  Statistique non-paramétrique dans un cadre général
  ESAIM:PS '13, SJS '14 , EJS '16 , Comm in Stat '18

  thèse

- Estimation de fonctionnelles liées à des croisements

  TEST '15, JSPI '19

Statistique des PDMPs

- Reconstruction des trajectoires ($\simeq$ estimation de Φ)

  Méthodes heuristiques pour des problèmes appliqués
  JRR '16, thèse de Florine Greciet

- Estimation des caractéristiques probabilistes (λ et Q)

  Statistique non-paramétrique dans un cadre général
  ESAIM:PS '13, SJS '14, EJS '16, Comm in Stat '18

  thèse

- Estimation de fonctionnelles liées à des croisements

  TEST '15, JSPI '19

Estimation du taux de saut λ : cadre de travail (1/2)

Observation parfaite ($\simeq \Phi$ connu) d'une seule trajectoire en temps long

- La loi conditionnelle est directement observée
- Absence de données i.i.d.
- Statistique asymptotique

Estimation du taux de saut λ : cadre de travail (1/2)

Observation parfaite ($\simeq \Phi$ connu) d'une seule trajectoire en temps long

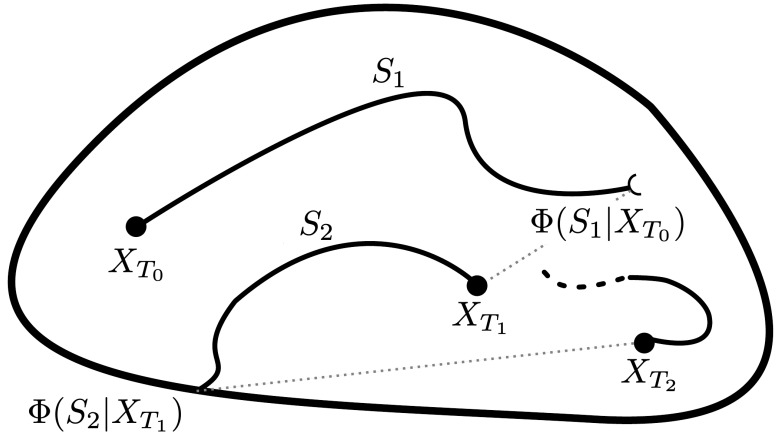
- La loi conditionnelle est directement observée
- Absence de données i.i.d.
- Statistique asymptotique

Hypothèses sur le modèle :

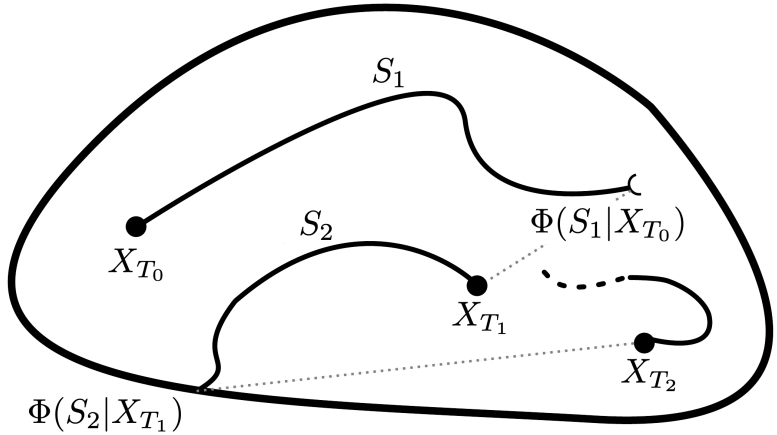
- La forme du modèle n'est pas spécifiée
- Hypothèse d'ergodicité pour garantir la qualité de l'estimation
- Espace d'état général

  Espace d'état typique : $E = \bigcup_{m \in M} \{m\} \times E_m$, avec $M \subset \mathbb{N}$ et $E_m \subset \mathbb{R}^{d_m}$

Estimation du taux de saut λ : cadre de travail (2/2)

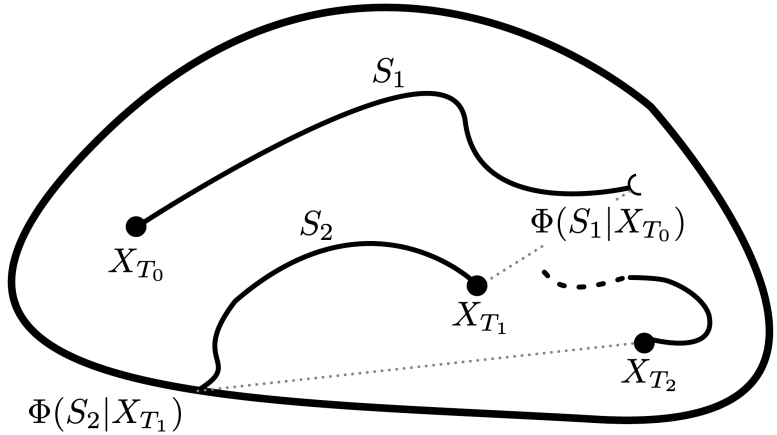


Estimation du taux de saut λ : cadre de travail (2/2)



$$\mathbb{P}(S_{n+1} > t | X_{T_n} = x) = \exp\left(-\int_0^t \lambda(\Phi(s|x))\mathrm{d}s\right) \mathbb{1}_{\{0 \leq t < t^+(x)\}}$$

Estimation du taux de saut λ : cadre de travail (2/2)



$$\mathbb{P}(S_{n+1} > t | X_{T_n} = x) = \underbrace{\exp\left(-\int_0^t \lambda(\Phi(s|x))\mathrm{d}s\right)}_{G(t|x)} \mathbb{1}_{\{0 \leq t < t^+(x)\}}$$

- Densité : $f(t|x) = -\frac{\partial G}{\partial t}(t|x)$
- Taux : $\lambda \circ \Phi(t|x) = \frac{f(t|x)}{G(t|x)}$

Stratégie 1

Appliquer le modèle à intensité multiplicative

SJS '14 et Comm in Stat '18

Estimateur de Nelson-Aalen

- Estimation directe d'un taux de saut (via lissage à noyau)
- Technique adaptée aux données censurées

Estimateur de Nelson-Aalen

- Estimation directe d'un taux de saut (via lissage à noyau)
- Technique adaptée aux données censurées
- Hypothèse : modèle à intensité multiplicative

Estimateur de Nelson-Aalen

- Estimation directe d'un taux de saut (via lissage à noyau)
- Technique adaptée aux données censurées
- Hypothèse : modèle à intensité multiplicative

Théorème (SJS '14)

- La plupart des PDMPs ne vérifient pas l'hypothèse d'intensité multiplicative

Estimateur de Nelson-Aalen

- Estimation directe d'un taux de saut (via lissage à noyau)
- Technique adaptée aux données censurées
- Hypothèse : modèle à intensité multiplicative

Théorème (SJS '14)

- La plupart des PDMPs ne vérifient pas l'hypothèse d'intensité multiplicative
- On peut estimer le taux d'un processus auxiliaire bien choisi :

$$\widetilde{\lambda}(t|x,y) = f(t|x)Q(y|\Phi(t|x))/H(y,t|x)$$

avec

$$H(y,t|x) = \int_t^{t^+(x)} f(t|x)Q(y|\Phi(t|x)) + G(t^+(x)|x)Q(y|\Phi(t^+(x)|x))$$

Retour au processus d'intérêt

- SJS '14 :

$$f(t|x) = \int_E \widetilde{\lambda}(t|x,y) H(y,t|x) \mathrm{d}y$$

Retour au processus d'intérêt

- SJS '14 :

$$f(t|x) = \int_E \widetilde{\lambda}(t|x,y) H(y,t|x) \mathrm{d}y$$

- Comm in Stat '18 :

$$\lambda(x) = \int_E \widetilde{\lambda}(0|x,y) R(y|x) \mathrm{d}y$$

où R est le noyau de (X_{T_n})

  - Décomposition de $\widetilde{\lambda}(\cdot\, t^+(x)|x,y)$ sur une base de $\mathbb{L}^2_{[0,1]}$
  - Estimation dans le cas discret : supp $Q(\mathrm{d}y|x)$ supposé fini et indépendant de x

Stratégie 2

Estimer $\lambda(\Phi(t|x))$ comme $f(t|x)/G(t|x)$

EJS '16

Estimateurs à noyau : définition

$$
\begin{aligned}
\widehat{\mathcal{F}}_n(x,t) &= \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{v_i^d w_i}\mathbb{K}_d\left(\frac{X_{T_i}-x}{v_i}\right)\mathbb{K}_1\left(\frac{S_{i+1}-t}{w_i}\right)\\
\widehat{\mathcal{G}}_n(x,t) &= \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{v_i^d}\mathbb{K}_d\left(\frac{X_{T_i}-x}{v_i}\right)\mathbb{1}_{\{S_{i+1}>t\}}\\
\widehat{\pi}_{\infty,n}(x) &= \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{v_i^d}\mathbb{K}_d\left(\frac{X_{T_i}-x}{v_i}\right)
\end{aligned}
$$

Fenêtres : $v_k = v_0(k+1)^{-\alpha}$ et $w_k = w_0(k+1)^{-\beta}$ avec $\alpha,\, \beta > 0$

$\mathbb{K}_p$ est un noyau sur $\mathbb{R}^p$, $p \in \{1, d\}$

Estimateurs à noyau : convergence

Théorème (EJS '16)

Pour $x \in E$ et $0 \leq t < t^+(x)$,

$$\begin{bmatrix} \widehat{\mathcal{F}}_n(x,t) \\ \widehat{\mathcal{G}}_n(x,t) \\ \widehat{\pi}_{\infty,n}(x) \end{bmatrix} \xrightarrow{\text{p.s}} \begin{bmatrix} \pi_\infty(x) f(t|x) \\ \pi_\infty(x) G(t|x) \\ \pi_\infty(x) \end{bmatrix}$$

et

$$n^{\frac{1-\alpha d-\beta}{2}} \left(\begin{bmatrix} \widehat{\mathcal{F}}_n(x,t) \\ \widehat{\mathcal{G}}_n(x,t) \\ \widehat{\pi}_{\infty,n}(x) \end{bmatrix} - \begin{bmatrix} \pi_\infty(x) f(t|x) \\ \pi_\infty(x) G(t|x) \\ \pi_\infty(x) \end{bmatrix} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0_3, \Sigma(x,t,\alpha,\beta)),$$

où la matrice de covariance $\Sigma(x,t,\alpha,\beta)$ est dégénérée avec un seul terme non nul en position $(1,1)$

Estimation de la composée $\lambda \circ \Phi$

$$\widehat{\lambda \circ \Phi}_n(t|x) = \frac{\widehat{\mathcal{F}}_n(x,t)}{\widehat{\mathcal{G}}_n(x,t)}$$

Corollaire (EJS '16)

Pour $x \in E$ et $0 < t < t^+(x)$,

$$\widehat{\lambda \circ \Phi}_n(t|x) \xrightarrow{\text{p.s.}} \lambda \circ \Phi(t|x)$$

et

$$n^{\frac{1-\alpha d-\beta}{2}} \left(\widehat{\lambda \circ \Phi}_n(t|x) - \lambda \circ \Phi(t|x) \right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau_1^2\, \tau_d^2\, \lambda \circ \Phi(t|x)}{(1+\alpha d+\beta)\pi_\infty(x) G(t|x)}\right)$$

Estimation du taux de saut λ (1/2)

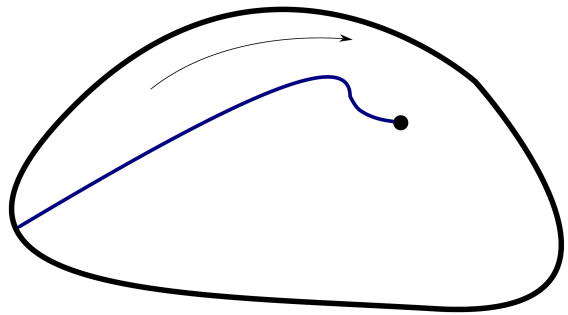
Cible : $\lambda(x)$ pour $x \in E$ fixé

Estimation du taux de saut λ (1/2)

Cible : $\lambda(x)$ pour $x \in E$ fixé

$$\mathcal{C}_x = \{\Phi(-t|x) \,:\, t \geq 0\} \cap E$$

$$\forall\, \xi \in \mathcal{C}_x,\ \exists!\, \tau_x(\xi),\ \Phi(\tau_x(\xi)|\xi) = x$$

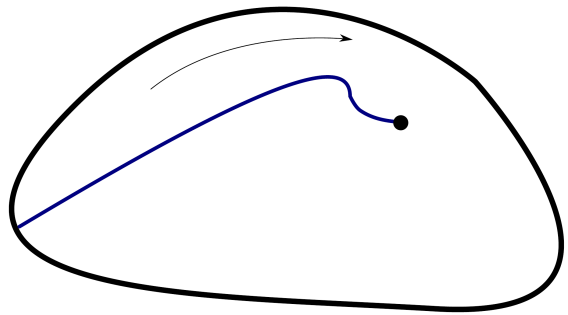


Estimation du taux de saut λ (1/2)

Cible : $\lambda(x)$ pour $x \in E$ fixé

$$\mathcal{C}_x = \{\Phi(-t|x) \,:\, t \geq 0\} \cap E$$

$$\forall\, \xi \in \mathcal{C}_x,\ \exists!\, \tau_x(\xi),\ \Phi(\tau_x(\xi)|\xi) = x$$



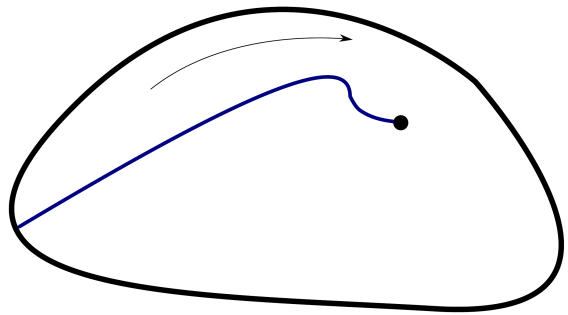
$$\lambda \circ \Phi(\tau_x(\xi)|\xi) = \lambda(x)$$

Estimation du taux de saut λ (1/2)

Cible : $\lambda(x)$ pour $x \in E$ fixé

$$\mathcal{C}_x = \{\Phi(-t|x) \,:\, t \geq 0\} \cap E$$

$$\forall\, \xi \in \mathcal{C}_x,\ \exists!\, \tau_x(\xi),\ \Phi(\tau_x(\xi)|\xi) = x$$



$$\lambda \circ \Phi(\tau_x(\xi)|\xi) = \lambda(x)$$

Pour tout $\xi \in \mathcal{C}_x$, $\boxed{\widehat{\lambda}_{\xi,n}(x) = \widehat{\lambda \circ \Phi}_n(\tau_x(\xi)|\xi) \text{ estime } \lambda(x)}$

Estimation du taux de saut λ (2/2)

Estimateur de variance asymptotique minimale :

$$\frac{\tau_1^2\,\tau_d^2\,\lambda \circ \Phi(\tau_x(\xi)|\xi)}{(1+\alpha d+\beta)\pi_\infty(\xi)G(t|\xi)} \propto (\pi_\infty(\xi)G(\tau_x(\xi)|\xi))^{-1}$$

$$\widehat{\lambda}_n(x) = \widehat{\lambda}_{\xi^\star,n}(x) \qquad \text{où} \quad \xi^\star = \arg\max_{\xi\in\mathcal{C}_x} \pi_\infty(\xi)G(\tau_x(\xi)|\xi)$$

Estimation du taux de saut λ (2/2)

Estimateur de variance asymptotique minimale :

$$\frac{\tau_1^2\,\tau_d^2\,\lambda \circ \Phi(\tau_x(\xi)|\xi)}{(1+\alpha d+\beta)\pi_\infty(\xi)G(t|\xi)} \propto (\pi_\infty(\xi)G(\tau_x(\xi)|\xi))^{-1}$$

$$\widehat{\lambda}_n(x) = \widehat{\lambda}_{\xi^\star,n}(x) \qquad \text{où} \quad \xi^\star = \arg\max_{\xi\in\mathcal{C}_x} \pi_\infty(\xi)G(\tau_x(\xi)|\xi)$$

$$\boxed{\widehat{\widehat{\lambda}}_n(x) = \widehat{\lambda}_{\xi^\star,n}(x) \qquad \text{où} \quad \xi^\star = \arg\max_{\xi\in\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi,\tau_x(\xi))}$$

Choix des paramètres de lissage

Critère ISE le long de $\mathcal{C}_x$:

$$\mathsf{ISE}_n(\alpha) = \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))^2 \mathrm{d}\xi - 2 \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))\, \pi_\infty(\xi) G(\tau_x(\xi)|\xi)\, \mathrm{d}\xi$$

Choix des paramètres de lissage

Critère ISE le long de $\mathcal{C}_x$:

$$\mathsf{ISE}_n(\alpha) = \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))^2 \mathrm{d}\xi - 2 \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))\, \pi_\infty(\xi) G(\tau_x(\xi)|\xi)\, \mathrm{d}\xi$$

Validation-croisée en dimension $d = 1$

$$\frac{1}{\overline{n}} \sum_{k=0}^{\overline{n}-1} \widehat{\mathcal{G}}_n(\overline{X}_{T_k}, \tau_x(\overline{X}_{T_k}))$$

$$\xrightarrow{\text{p.s.}} \int \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))\, \pi_\infty(\xi) \qquad \mathrm{d}\xi,$$

où $(\overline{X}_{T_k}, \overline{S}_{k+1})$ est la chaîne immergée d'un autre PDMP, indépendant du premier, et généré selon les mêmes paramètres

Choix des paramètres de lissage

Critère ISE le long de $\mathcal{C}_x$:

$$\mathsf{ISE}_n(\alpha) = \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))^2 \mathrm{d}\xi - 2 \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))\, \pi_\infty(\xi) G(\tau_x(\xi)|\xi)\, \mathrm{d}\xi$$

Validation-croisée en dimension $d = 1$

$$\frac{1}{\overline{n}} \sum_{k=0}^{\overline{n}-1} \widehat{\mathcal{G}}_n(\overline{X}_{T_k}, \tau_x(\overline{X}_{T_k}))\, \mathbb{1}_{\mathcal{C}_x}(\overline{X}_{T_k})$$

$$\xrightarrow{\text{p.s.}} \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))\, \pi_\infty(\xi) \qquad \mathrm{d}\xi,$$

où $(\overline{X}_{T_k}, \overline{S}_{k+1})$ est la chaîne immergée d'un autre PDMP, indépendant du premier, et généré selon les mêmes paramètres

Choix des paramètres de lissage

Critère ISE le long de $\mathcal{C}_x$:

$$\mathsf{ISE}_n(\alpha) = \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))^2 \mathrm{d}\xi - 2 \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))\, \pi_\infty(\xi) G(\tau_x(\xi)|\xi)\, \mathrm{d}\xi$$

Validation-croisée en dimension $d = 1$

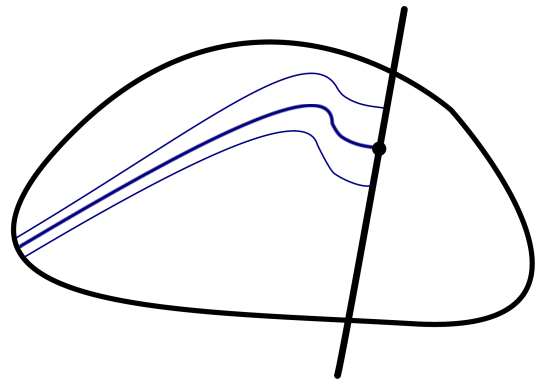
$$\frac{1}{\overline{n}} \sum_{k=0}^{\overline{n}-1} \widehat{\mathcal{G}}_n(\overline{X}_{T_k}, \tau_x(\overline{X}_{T_k}))\, \mathbb{1}_{\mathcal{C}_x}(\overline{X}_{T_k}) \mathbb{1}_{(\tau_x(\overline{X}_{T_k}),\infty)}(\overline{S}_{k+1})$$

$$\xrightarrow{\text{p.s.}} \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))\, \pi_\infty(\xi) G(\tau_x(\xi)|\xi)\, \mathrm{d}\xi,$$

où $(\overline{X}_{T_k}, \overline{S}_{k+1})$ est la chaîne immergée d'un autre PDMP, indépendant du premier, et généré selon les mêmes paramètres

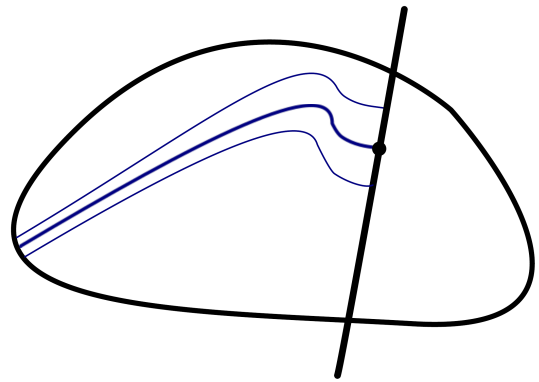
Validation-croisée en dimension $d > 1$

- $\mathbb{H}_x = \left\{ y \in \mathbb{R}^d \ : \ y - x \perp \dfrac{\partial \Phi}{\partial t}(0|x) \right\}$
- $\mathbb{D}_{x,\rho} = B_d(x, \rho) \cap \mathbb{H}_x$
- $\mathbb{T}_{x,\rho} = \bigcup_{y \in \mathbb{D}_{x,\rho}} \mathcal{C}_y$



Validation-croisée en dimension $d > 1$

- $\mathbb{H}_x = \left\{ y \in \mathbb{R}^d \ : \ y - x \perp \dfrac{\partial \Phi}{\partial t}(0|x) \right\}$
- $\mathbb{D}_{x,\rho} = B_d(x,\rho) \cap \mathbb{H}_x$
- $\mathbb{T}_{x,\rho} = \bigcup_{y \in \mathbb{D}_{x,\rho}} \mathcal{C}_y$



Approximation du critère ISE :

$$\widehat{\mathsf{ISE}}_{n,\overline{n}}(\alpha) \ = \ \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_n(\xi, \tau_x(\xi))^2 \mathrm{d}\xi$$
$$- \ \frac{2\,\Gamma\left(\frac{d-1}{2}+1\right)}{\overline{n}\,\pi^{\frac{d-1}{2}}\,\rho^{d-1}} \sum_{k=0}^{\overline{n}-1} \widehat{\mathcal{G}}_n\big(\overline{X}_{T_k}, \tau_x(\overline{X}_{T_k})\big)\, \mathbb{1}_{\mathbb{T}_{x,\rho}}(\overline{X}_{T_k})\, \mathbb{1}_{(\tau_x(\overline{X}_{T_k}),\infty)}(\overline{S}_{k+1})$$

Proposition (EJS '16)

Conditionnellement à l'observation de (X_t), quand $\overline{n} \to \infty$ et $\rho \to 0$,

$$\widehat{\mathsf{ISE}}_{n,\overline{n}}(\alpha) \xrightarrow{\text{p.s.}} \mathsf{ISE}_n(\alpha)$$

Estimation du taux de saut λ : perspectives (1/2)

Application du modèle à intensité multiplicative

Problème principal : revenir au processus d'intérêt, i.e. inverser

$$\widetilde{\lambda}(t|x,y) = f(t|x)Q(y|\Phi(t|x))/H(y,t|x)$$

Estimation du taux de saut λ : perspectives (1/2)

Application du modèle à intensité multiplicative

Problème principal : revenir au processus d'intérêt, i.e. inverser

$$\widetilde{\lambda}(t|x,y) = f(t|x)Q(y|\Phi(t|x))/H(y,t|x)$$

- Comm in Stat '18 : supp $Q(\mathrm{d}y|x)$ supposé fini et indépendant de x

  $\rightarrow$ Quantifier la loi invariante π_∞ (ou quantifier R ?)

- Estimer $\widetilde{f}(t|x,y)$ (Kaplan-Meier) plutôt que $\widetilde{\lambda}(t|x,y)$ puis utiliser

$$\widetilde{f}(t|x,y) \propto f(t|x)Q(y|\Phi(t|x))$$

Estimation du taux de saut λ : perspectives (1/2)

Application du modèle à intensité multiplicative

Problème principal : revenir au processus d'intérêt, i.e. inverser

$$\widetilde{\lambda}(t|x,y) = f(t|x)Q(y|\Phi(t|x))/H(y,t|x)$$

- Comm in Stat '18 : supp $Q(\mathrm{d}y|x)$ supposé fini et indépendant de x

  $\rightarrow$ Quantifier la loi invariante π_∞ (ou quantifier R ?)

- Estimer $\widetilde{f}(t|x,y)$ (Kaplan-Meier) plutôt que $\widetilde{\lambda}(t|x,y)$ puis utiliser

$$\widetilde{f}(t|x,y) \propto f(t|x)Q(y|\Phi(t|x))$$

- Lissage spatial (uniquement en x ?) pour faciliter la comparaison avec l'estimateur quotient

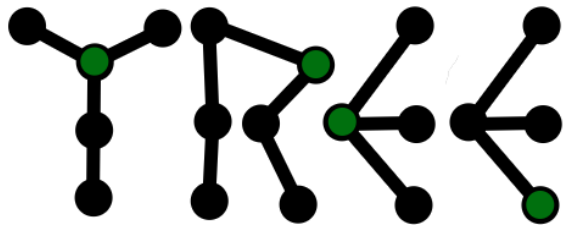
Estimation du taux de saut λ : perspectives (2/2)

Estimateur quotient

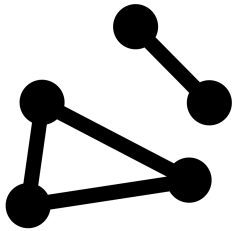
- Convergence de $\widehat{\widehat{\lambda}}_n(x)$
- Procédure d'estimation ponctuelle $\rightarrow$ procédure d'estimation globale ?
- Aggrégation d'estimateurs :

$$\widehat{\lambda}_n(x) = \int_{\mathcal{C}_x} \alpha(\xi)\widehat{\lambda}_{\xi,n}(x)\mathrm{d}\xi \qquad \text{où} \int_{\mathcal{C}_x} \alpha(\xi)\mathrm{d}\xi = 1$$

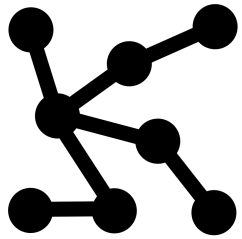
$\rightarrow$ Étude des corrélations des $\widehat{\lambda}_{\xi,n}(x)$


TREE

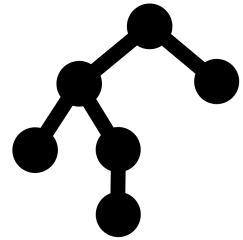
Arbres enracinés non-ordonnés



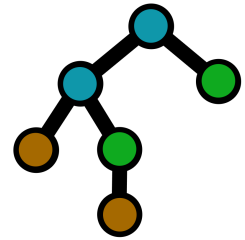
graphe



arbre

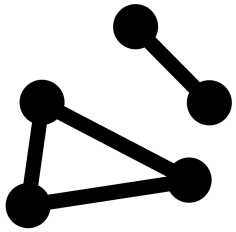


arbre enraciné

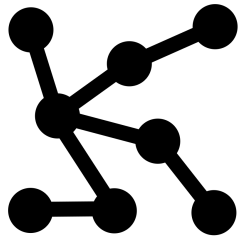


arbre enraciné étiqueté

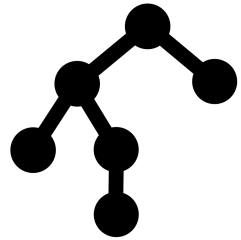
Arbres enracinés non-ordonnés



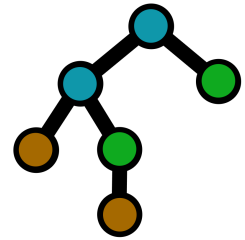
graphe



arbre



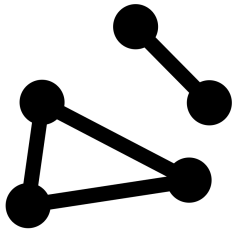
arbre enraciné



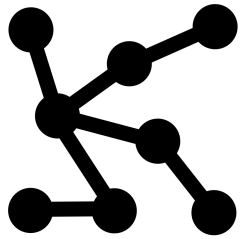
arbre enraciné étiqueté

crédit photo : Benoît Henry

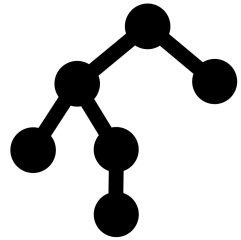
Arbres enracinés non-ordonnés



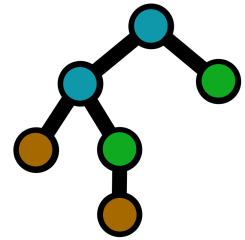
graphe



arbre



arbre enraciné



arbre enraciné étiqueté

Modélisation de

- structures aériennes de plantes
- lignées cellulaires
- fichiers `XML`

séquences $\subset$ arbres $\subset$ graphes

crédit photo : Benoît Henry

Statistique des données arborescentes

- Modèle stochastique génératif

- Étude de la distribution via la détection de motifs

Statistique des données arborescentes

- Modèle stochastique génératif

  Processus de Galton-Watson
  ALEA '19, Preprint

- Étude de la distribution via la détection de motifs

Statistique des données arborescentes

- Modèle stochastique génératif

  Processus de Galton-Watson
  ALEA '19, Preprint

- Étude de la distribution via la détection de motifs de type sous-arbre :
  - Motifs exacts et noyaux de convolution

    JMLR '20 , TCS '22 , thèse de Florian Ingels
  - Motifs approchés

Statistique des données arborescentes

- Modèle stochastique génératif

  Processus de Galton-Watson
  ALEA '19, Preprint

- Étude de la distribution via la détection de motifs de type sous-arbre :
  - Motifs exacts et noyaux de convolution

    JMLR '20 , TCS '22 , thèse de Florian Ingels
  - Motifs approchés :
    - Approximation topologique
      Algorithms '19, ALENEX '19
    - Approximation des étiquettes
      IWOCA '21
    - Approximation topologique et des étiquettes
      FSPM '23

Statistique des données arborescentes

- Modèle stochastique génératif

  Processus de Galton-Watson
  ALEA '19, Preprint

- Étude de la distribution via la détection de motifs de type sous-arbre :
  - Motifs exacts et noyaux de convolution
    JMLR '20, TCS '22, thèse de Florian Ingels
  - Motifs approchés :
    - Approximation topologique
      Algorithms '19, ALENEX '19
    - Approximation des étiquettes
      IWOCA '21
    - Approximation topologique et des étiquettes
      FSPM '23

Autour du noyau des sous-arbres

JMLR '20 et TCS '22

Noyaux

- $K : \mathcal{X}^2 \to \mathbb{R}$ est un noyau sur $\mathcal{X}$
  si les matrices $[K(X_i, X_j)]_{i,j}$ sont symétriques semi-définies positives
- Moore-Aronszajn : $K(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$
- Interface avec des algorithmes efficaces, e.g. SVM, ACP, NN

Noyaux

- $K : \mathcal{X}^2 \to \mathbb{R}$ est un noyau sur $\mathcal{X}$
  si les matrices $[K(X_i, X_j)]_{i,j}$ sont symétriques semi-définies positives
- Moore-Aronszajn : $K(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$
- Interface avec des algorithmes efficaces, e.g. SVM, ACP, NN

Noyaux sur des structures combinatoires : convolution (Haussler '99)

$$K(x, y) = \sum_{s \in \mathcal{S}_{\mathcal{X}}} \omega_s \varphi(\mathsf{num}_s(x), \mathsf{num}_s(y)),$$

où

- $\mathcal{S}_{\mathcal{X}}$ désigne un ensemble de sous-structures des éléments de $\mathcal{X}$
- $\mathsf{num}_s(x)$ est le nombre d'occurences de s dans x
- ω_s est le poids associé à s
- φ est un noyau sur $\mathbb{N}$ ou sur $\mathbb{R}$

Noyaux

- $K : \mathcal{X}^2 \to \mathbb{R}$ est un noyau sur $\mathcal{X}$
  si les matrices $[K(X_i, X_j)]_{i,j}$ sont symétriques semi-définies positives
- Moore-Aronszajn : $K(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$
- Interface avec des algorithmes efficaces, e.g. SVM, ACP, NN

Noyaux sur des structures combinatoires : convolution (Haussler '99)

$$K(x, y) = \sum_{s \in \mathcal{S}_x \cap \mathcal{S}_y} \omega_s \, \mathrm{num}_s(x) \, \mathrm{num}_s(y) \, ,$$

où

- $\mathcal{S}_{\mathcal{X}}$ désigne un ensemble de sous-structures des éléments de $\mathcal{X}$
- $\mathrm{num}_s(x)$ est le nombre d'occurences de s dans x
- ω_s est le poids associé à s
- $\varphi(x, y) = xy$

Sélection de l'ensemble de sous-structures

Ensemble de sous-structures riche, e.g. sous-graphes pour les graphes

- \+ Comparaison pertinente
- − Complexité temporelle mauvaise

Ensemble de sous-structures pauvre, e.g. lettres pour les séquences

- − Comparaison naïve
- \+ Complexité temporelle rapide

Sélection de l'ensemble de sous-structures

Ensemble de sous-structures riche, e.g. sous-graphes pour les graphes

- \+ Comparaison pertinente
- − Complexité temporelle mauvaise

Ensemble de sous-structures pauvre, e.g. lettres pour les séquences

- − Comparaison naïve
- \+ Complexité temporelle rapide

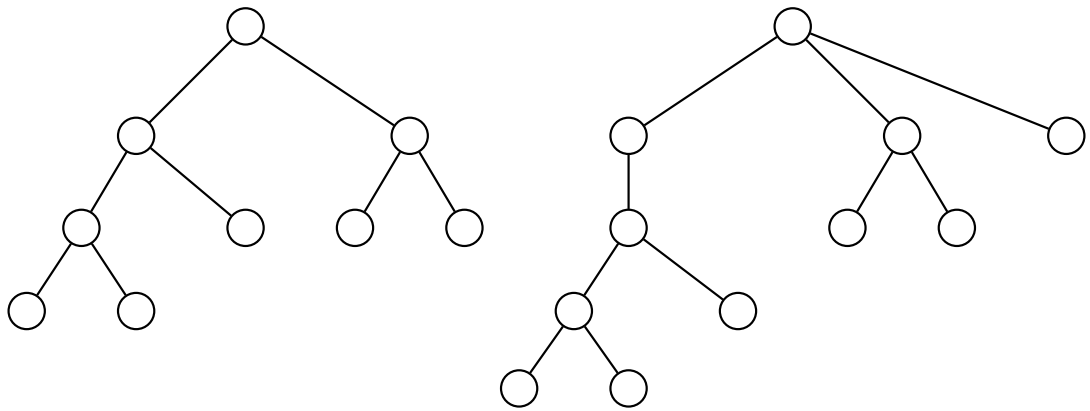
Évaluation de la matrice de Gram $[K(X_i, X_j)]_{i,j}$

- Évaluation directe de $\text{num}_s(X_i)$ et $\text{num}_s(X_j)$ pour $s \in \mathcal{S}_{X_i} \cap \mathcal{S}_{X_j}$
- Énumération préliminaire de $\bigcup_{1 \leq i \leq n} \mathcal{S}_{X_i}$

Le noyau des sous-arbres (Vishwanathan et Smola '02)

$$K(x,y) = \sum_{s \in \mathcal{S}_x \cap \mathcal{S}_y} \omega_s \, \mathsf{num}_s(x) \, \mathsf{num}_s(y)$$

T_1 T_2

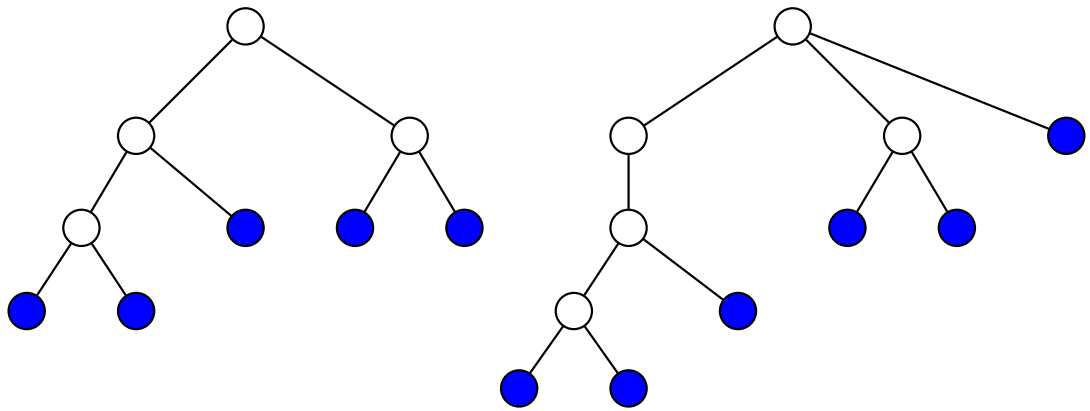


$K(T_1, T_2) =$

Le noyau des sous-arbres (Vishwanathan et Smola '02)

$$K(x,y) = \sum_{s \in \mathcal{S}_x \cap \mathcal{S}_y} \omega_s \, \mathsf{num}_s(x) \, \mathsf{num}_s(y)$$

T_1 T_2

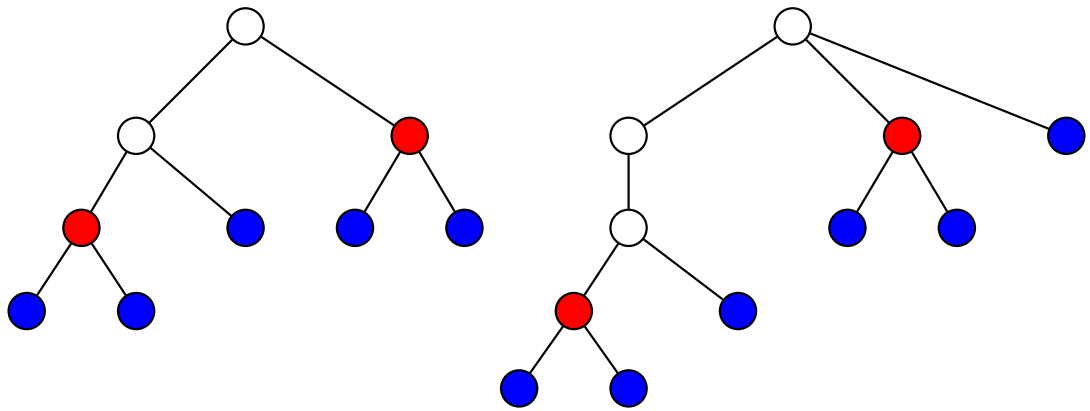


$K(T_1, T_2) = \omega_{\bullet} \times 5 \times 6$

Le noyau des sous-arbres (Vishwanathan et Smola '02)

$$K(x,y) = \sum_{s \in \mathcal{S}_x \cap \mathcal{S}_y} \omega_s \, \mathsf{num}_s(x) \, \mathsf{num}_s(y)$$

T_1 T_2

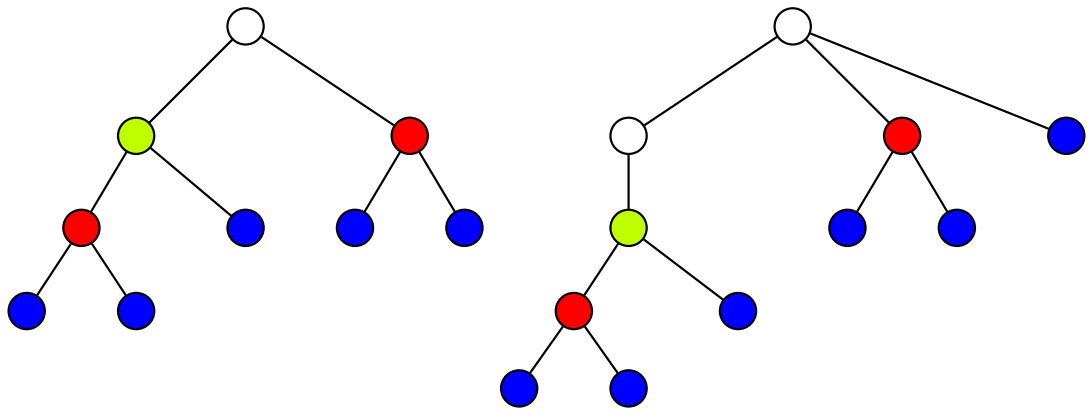


$$K(T_1, T_2) = \omega_{\bullet} \times 5 \times 6 + \omega_{\bullet} \times 2 \times 2$$

Le noyau des sous-arbres (Vishwanathan et Smola '02)

$$K(x,y) = \sum_{s \in \mathcal{S}_x \cap \mathcal{S}_y} \omega_s \, \mathsf{num}_s(x) \, \mathsf{num}_s(y)$$

T_1 T_2

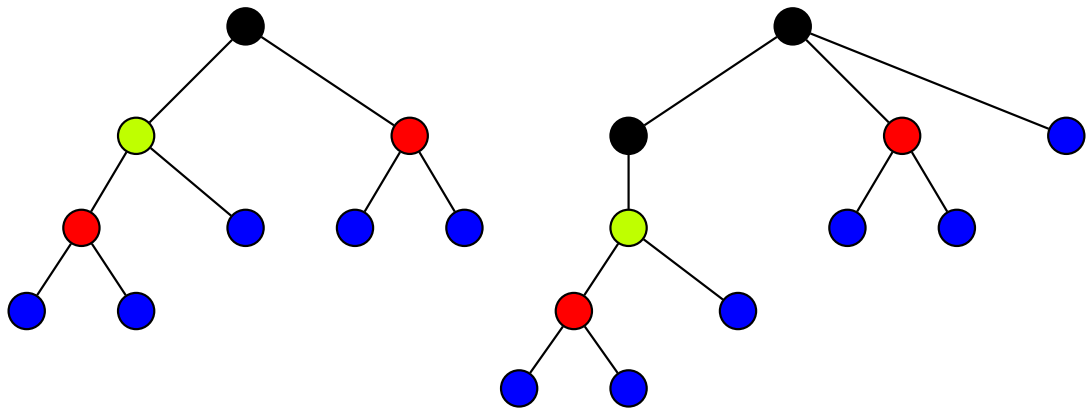


$$K(T_1, T_2) = \omega_{\bullet} \times 5 \times 6 + \omega_{\bullet} \times 2 \times 2 + \omega_{\bullet} \times 1 \times 1$$

Le noyau des sous-arbres (Vishwanathan et Smola '02)

$$K(x,y) = \sum_{s \in \mathcal{S}_x \cap \mathcal{S}_y} \omega_s \, \mathsf{num}_s(x) \, \mathsf{num}_s(y)$$

T_1 T_2

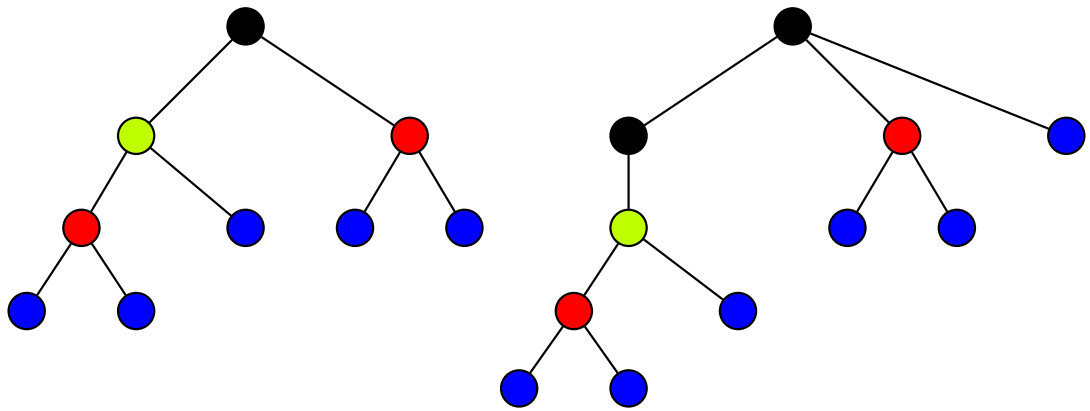


$$K(T_1, T_2) = \omega_{\bullet} \times 5 \times 6 + \omega_{\bullet} \times 2 \times 2 + \omega_{\bullet} \times 1 \times 1$$

Le noyau des sous-arbres (Vishwanathan et Smola '02)

$$K(x,y) = \sum_{s \in \mathcal{S}_x \cap \mathcal{S}_y} \omega_s \, \mathsf{num}_s(x) \, \mathsf{num}_s(y)$$

T_1 T_2



$$K(T_1, T_2) = \omega_{\bullet} \times 5 \times 6 + \omega_{\bullet} \times 2 \times 2 + \omega_{\bullet} \times 1 \times 1$$

Algorithme : formule récursive pour $K(T_1, T_2)$ en $O(\#T_1 + \#T_2)$

Hypothèse : $\omega_s = \lambda^{\#s}$ où $\lambda^{\mathsf{height}(s)}$, $0 < \lambda \leq 1$

Tentative d'étude théorique du noyau des sous-arbres (1/2)

- Modèle stochastique à 2 classes
- ρ : paramètre de dissimilarité des classes
- Classifieur : $T \mapsto \arg\max_k$ mean $\{K(T,t) : \text{class}(t) = k\}$

Tentative d'étude théorique du noyau des sous-arbres (1/2)

- Modèle stochastique à 2 classes
- ρ : paramètre de dissimilarité des classes
- Classifieur : $T \mapsto \arg\max_k$ mean $\{K(T,t) : \text{class}(t) = k\}$

Proposition (JMLR '20)

Un jeu de données d'apprentissage de taille

$$\frac{2\max_k K(T_k,T_k)^2}{\min_k C_{k,h}^2}\frac{\exp(2\rho)}{H^2}\log\left(\frac{2}{\delta}\right)$$

est suffisant pour que, avec probabilité au moins $1-\delta$, le classifieur induise une erreur d'au plus $1 - F_\rho(h) + \delta$

- $C_{k,h} = \dfrac{K(T_k,T_k) - \max\limits_{\{u \in T_k \,:\, \text{height}(T_k[u])=h\}} K(T_k[u],T_k[u])}{\#\text{leaves}(T_k)}$
- $F_\rho(h)$: fonction de répartition de la loi binomiale de paramètre $(H, \rho/H)$

Tentative d'étude théorique du noyau des sous-arbres (2/2)

Corollaire (JMLR '20)

Le nombre de données d'apprentissage susmentionné est minimal lorsque $\omega_{\mathsf{leaf}} = 0$

Tentative d'étude théorique du noyau des sous-arbres (2/2)

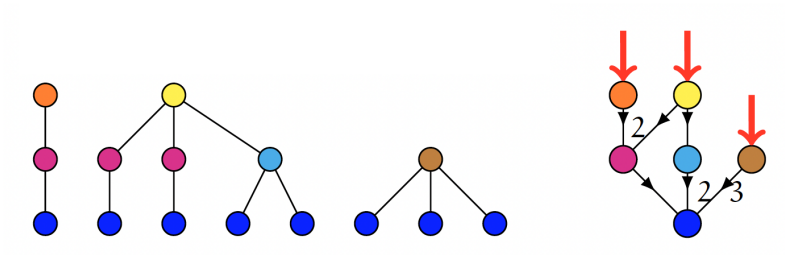
Corollaire (JMLR '20)

Le nombre de données d'apprentissage susmentionné est minimal lorsque $\omega_{\mathsf{leaf}} = 0$

Limite du calcul récursif :

- Avec $\omega_s = \lambda^{\#s}$ ou $\lambda^{\mathsf{height}(s)}$, le poids des feuilles est maximal
- Le poids devrait dépendre des fréquences d'apparition des sous-structures dans les classes

Énumération des sous-arbres : algorithme



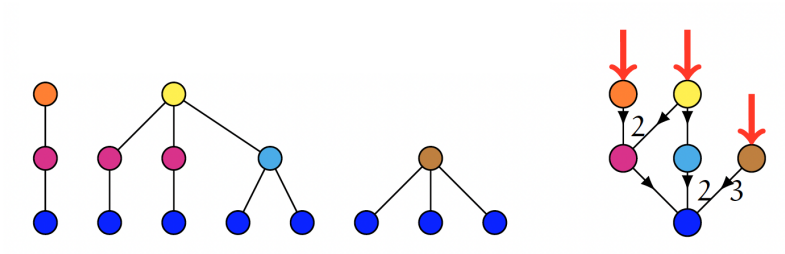
- Énumération préliminaire de $\bigcup_{1 \leq i \leq n} \mathcal{S}_{T_i}$ en temps quadratique

  ↑

  tous les sous-arbres des données

- Évaluation de $K(T_i, T_j)$ en $O(\min(\#T_i, \#T_j))$

Énumération des sous-arbres : algorithme



- Énumération préliminaire de $\bigcup_{1 \leq i \leq n} \mathcal{S}_{T_i}$ en temps quadratique

  $\uparrow$

  tous les sous-arbres des données

- Évaluation de $K(T_i, T_j)$ en $O(\min(\#T_i, \#T_j))$

Adapté à :

- Des évaluations répétées de la matrice de Gram, e.g. pour tester différentes valeurs de λ
- Apprendre la fonction de poids à partir des données

Énumération des sous-arbres : fonction de poids

$\rho_s(k) =$ fréquence de la sous-structure s dans la classe k

Conjecture

Si ρ_s est proche de $1_k = (0, \ldots, \underset{\uparrow}{1}, \ldots, 0)$ ou de $\overline{1}_k = (1, \ldots, \underset{\uparrow}{0}, \ldots, 1)$,

k $\qquad$ k

alors s aide à discriminer la classe k

Énumération des sous-arbres : fonction de poids

$\rho_s(k) =$ fréquence de la sous-structure s dans la classe k

Conjecture

Si ρ_s est proche de $1_k = (0, \ldots, \underset{\uparrow}{1}, \ldots, 0)$ ou de $\overline{1}_k = (1, \ldots, \underset{\uparrow}{0}, \ldots, 1)$,

$\qquad\qquad\qquad\qquad\qquad\qquad\quad k \qquad\qquad\qquad\qquad\qquad\qquad\quad k$

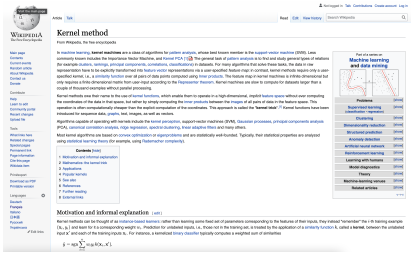
alors s aide à discriminer la classe k

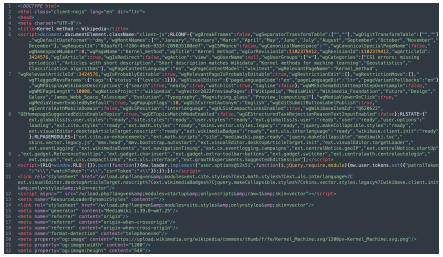
ω_s : fonction décroissante de la distance minimale aux 1_k's et aux $\overline{1}_k$'s

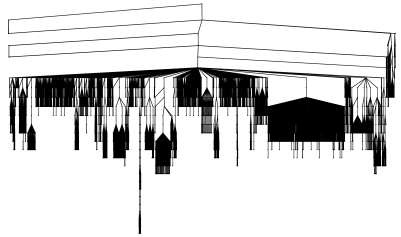
- si la distance est petite, alors s est pertinente et ω_s est grand
- si la distance est grande, alors s n'est pas pertinente et ω_s est petit

Énumération des sous-arbres : application à des données réelles

Langue d'un article Wikipédia à partir de la topologie de son code `HTML` ?

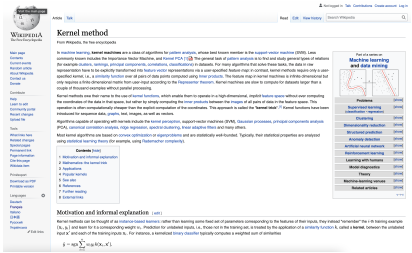


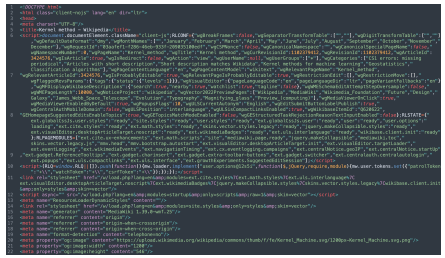


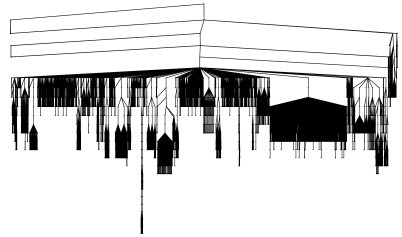


Énumération des sous-arbres : application à des données réelles

Langue d'un article Wikipédia à partir de la topologie de son code `HTML` ?

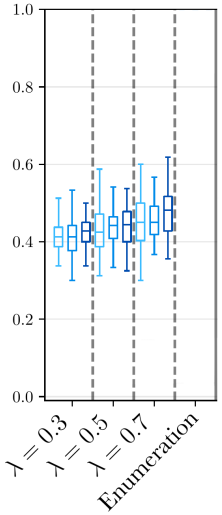






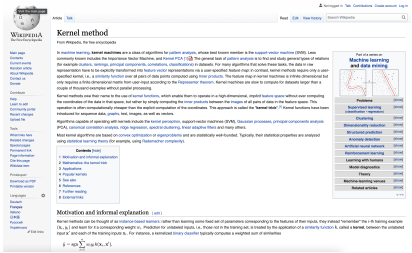
Noyau des sous-arbres feat. SVM

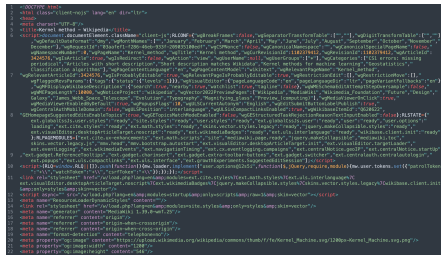
Précision

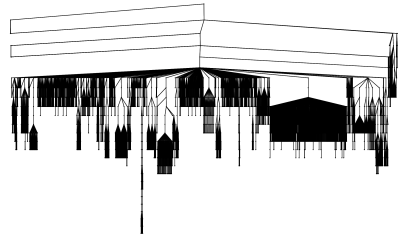


Énumération des sous-arbres : application à des données réelles

Langue d'un article Wikipédia à partir de la topologie de son code `HTML` ?

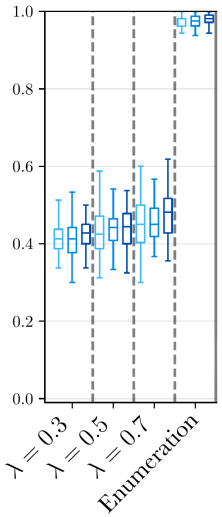






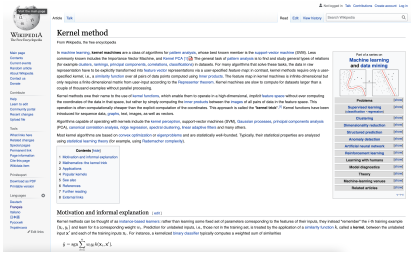
Noyau des sous-arbres feat. SVM

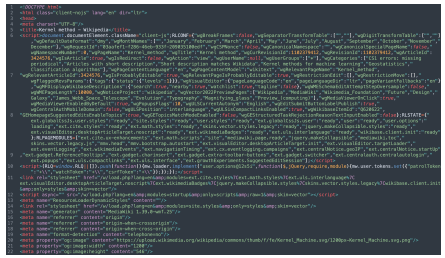
Précision

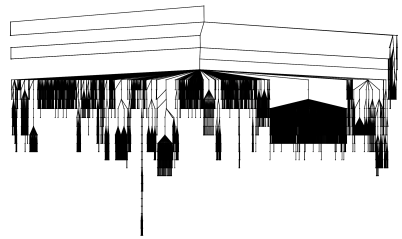


Énumération des sous-arbres : application à des données réelles

Langue d'un article Wikipédia à partir de la topologie de son code `HTML` ?

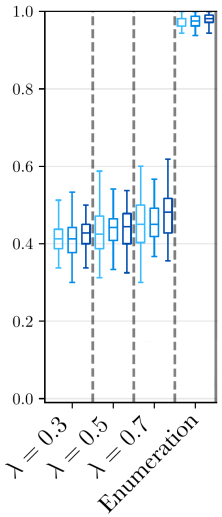




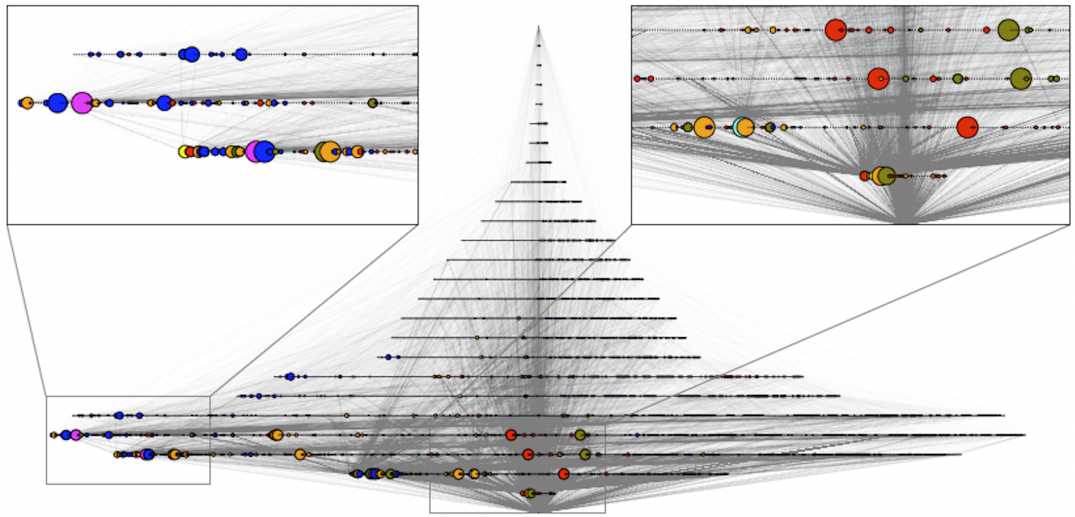


Noyau des sous-arbres feat. SVM

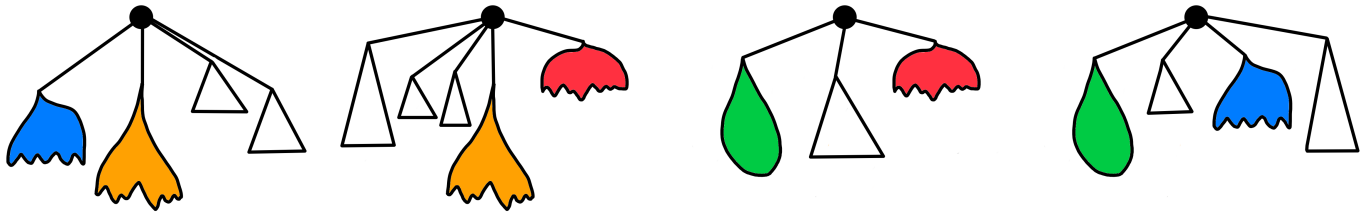
Précision



Visualisation des données

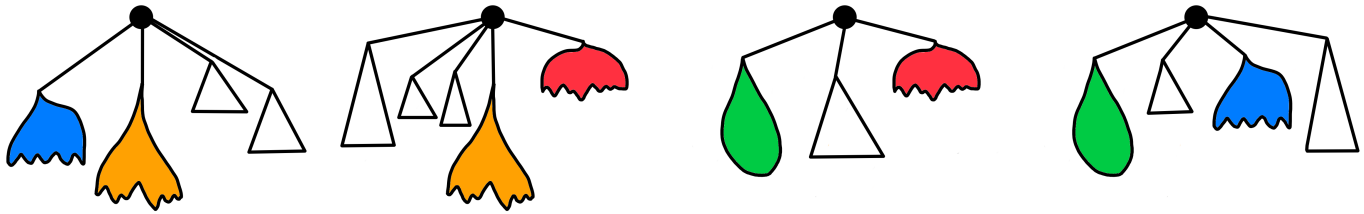


Limite du noyau des sous-arbres



Noyau des sous-arbres inefficace sur cet exemple

Limite du noyau des sous-arbres



Noyau des sous-arbres inefficace sur cet exemple

$\rightarrow$ Énumération des sous-forêts
$\quad \subset$ Énumération des forêts

$\rightarrow$ Même procédure d'apprentissage

Énumération des forêts non-redondantes (1/2)

$(T_1, \ldots, T_n)$ est une forêt non-redondante si, pour tout couple (i, j), T_i n'est pas un sous-arbre de T_j

Énumération des forêts non-redondantes (1/2)

$(T_1, \ldots, T_n)$ est une forêt non-redondante si, pour tout couple (i, j), T_i n'est pas un sous-arbre de T_j

Algorithme de recherche inversée sous forme compressée

Théorème (TCS '22)

L'algorithme construit un arbre d'énumération des forêts non-redondantes

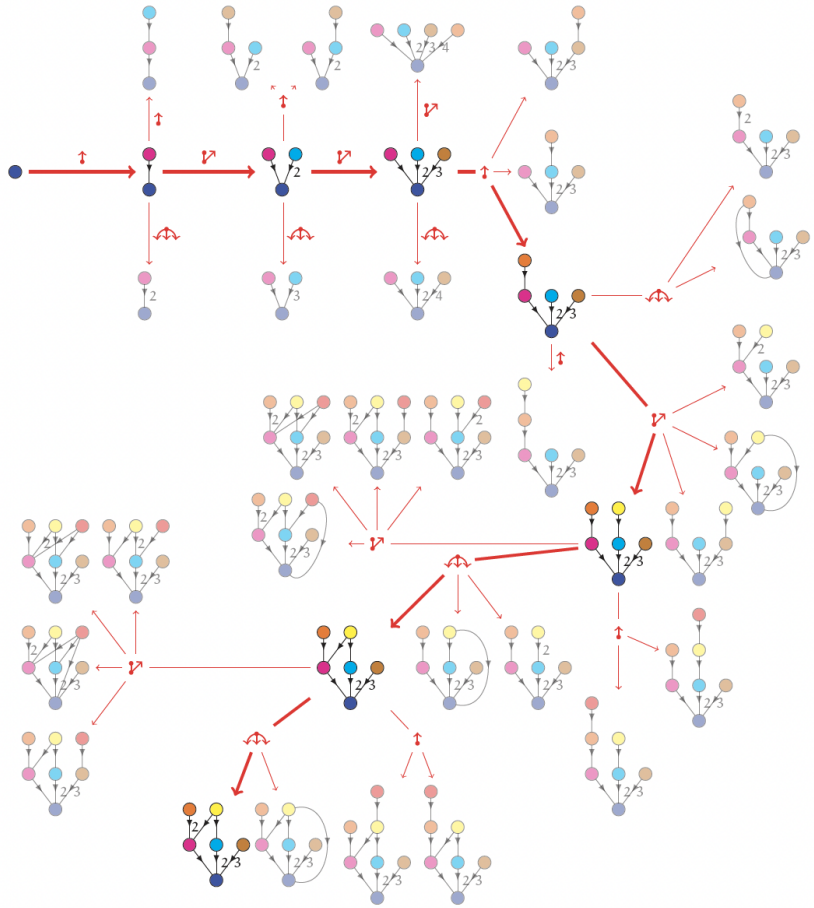
Énumération des forêts non-redondantes (1/2)

$(T_1, \ldots, T_n)$ est une forêt non-redondante si, pour tout couple (i, j), T_i n'est pas un sous-arbre de T_j

Algorithme de recherche inversée sous forme compressée

Théorème (TCS '22)

L'algorithme construit un arbre d'énumération des forêts non-redondantes



Énumération des forêts non-redondantes (2/2)

Contrôle de l'arbre d'énumération

Théorème (TCS '22)

- Nombre de successeurs d'une forêt Δ est $\Theta(\#\Delta)$
- Construction (incrémentale) en temps $O(\#\Delta \deg(\Delta))$

Énumération des forêts non-redondantes (2/2)

Contrôle de l'arbre d'énumération

Théorème (TCS '22)

- Nombre de successeurs d'une forêt Δ est $\Theta(\#\Delta)$
- Construction (incrémentale) en temps $O(\#\Delta \deg(\Delta))$

- E_K : forêts à profondeur K dans l'arbre d'énumération
- $E_{\leq K} = \bigcup_{0 \leq k \leq K} E_k$

Énumération des forêts non-redondantes (2/2)

Contrôle de l'arbre d'énumération

Théorème (TCS '22)

- Nombre de successeurs d'une forêt Δ est $\Theta(\#\Delta)$
- Construction (incrémentale) en temps $O(\#\Delta \deg(\Delta))$
- Énumérer $E_{\leq K+1}$ se fait en temps $O(K^2 \#E_{\leq K})$

- E_K : forêts à profondeur K dans l'arbre d'énumération
- $E_{\leq K} = \bigcup_{0 \leq k \leq K} E_k$

Énumération des forêts non-redondantes (2/2)

Contrôle de l'arbre d'énumération

Théorème (TCS '22)

- Nombre de successeurs d'une forêt Δ est $\Theta(\#\Delta)$
- Construction (incrémentale) en temps $O(\#\Delta \deg(\Delta))$
- Énumérer $E_{\leq K+1}$ se fait en temps $O(K^2 \#E_{\leq K})$
- Quand $K \to \infty$,

$$E_K = K! \left(\frac{12}{\pi^2}\right)^K \left(\frac{6\sqrt{2}}{\pi^2} \exp\left(\frac{\pi^2}{24}\right) + O\left(\frac{1}{K}\right)\right)$$

- E_K : forêts à profondeur K dans l'arbre d'énumération
- $E_{\leq K} = \bigcup_{0 \leq k \leq K} E_k$

Autour du noyau des sous-arbres : perspectives (1/3)

- Propriétés de la fonction de poids :
  - Problème à 2 classes

    Poids des feuilles nul
    $\rightarrow$ Poids des sous-structures communes nul
  - Problème multi-classes

    $\rightarrow$ Poids des sous-structures présentes ou absentes dans plusieurs classes nul
    $\rightarrow$ Poids fonction de la distance à ces cas extrêmes

Autour du noyau des sous-arbres : perspectives (1/3)

- Propriétés de la fonction de poids :
  - Problème à 2 classes

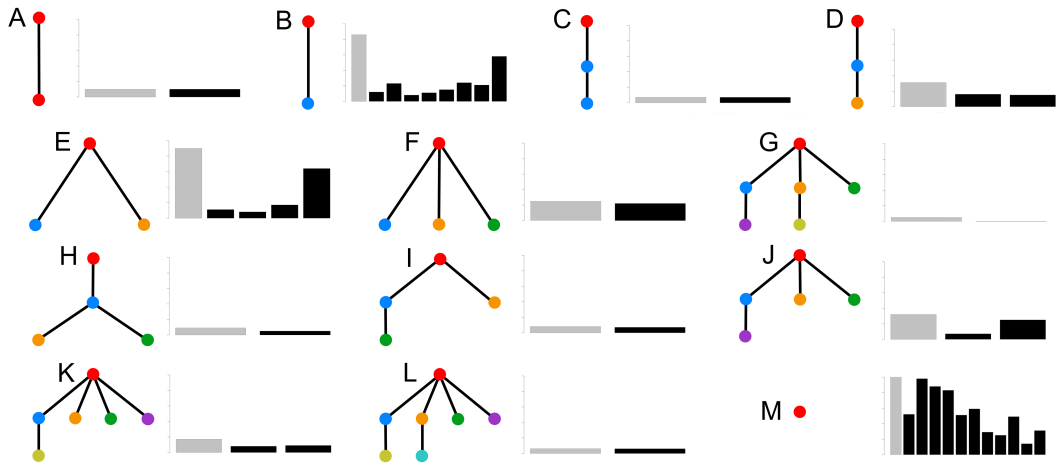
    Poids des feuilles nul
    $\rightarrow$ Poids des sous-structures communes nul
  - Problème multi-classes

    $\rightarrow$ Poids des sous-structures présentes ou absentes dans plusieurs classes nul
    $\rightarrow$ Poids fonction de la distance à ces cas extrêmes

- Noyau des sous-forêts :
  - Temps de calcul $\rightarrow$ restriction aux sous-forêts fréquentes
  - Compensation : méthode de Nyström ?

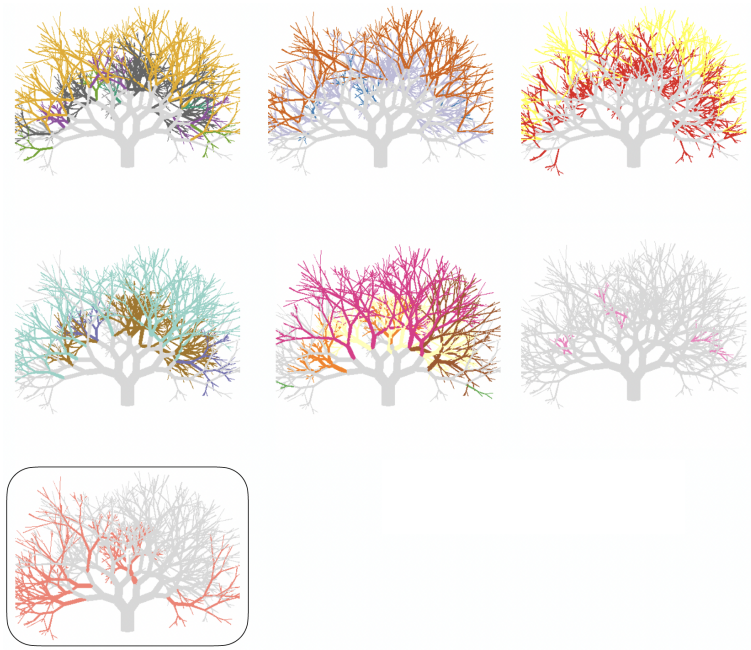
Autour du noyau des sous-arbres : perspectives (2/3)

- Sous-arbres avec distribution commune des étiquettes :
  - Décider si deux arbres isomorphes partagent la même distribution des étiquettes : GI-complet
  - Heuristique proposée dans IWOCA '21
  - Détection des sous-structures fréquentes :



Autour du noyau des sous-arbres : perspectives (3/3)

- Approximation topologique et géométrique des sous-arbres :
  - Clustering en distance d'édition
  - Clusters vs antichaînes
  - Suppression du bruit et visualisation condensée de l'arbre :



Élaboration d'un nouveau projet

Alternatives aux MCMC pour la simulation de modèles sur réseau

Mesure de Gibbs : $\mu_\beta(\sigma) \propto \exp(-\beta \mathcal{H}(\sigma))$

avec $\mathcal{H}(\sigma) = \sum_{\{i,j\} \in \mathcal{E}} \sigma_i \sigma_j - h \sum_{i \in \mathcal{V}} \sigma_i$

Simulation : MCMC

Élaboration d'un nouveau projet

Alternatives aux MCMC pour la simulation de modèles sur réseau

Mesure de Gibbs : $\mu_\beta(\sigma) \propto \exp(-\beta \mathcal{H}(\sigma))$

avec $\mathcal{H}(\sigma) = \sum_{\{i,j\} \in \mathcal{E}} \sigma_i \sigma_j - h \sum_{i \in \mathcal{V}} \sigma_i$

Simulation : MCMC vs nouvelles techniques

- Pirogov-Sinai algorithmique (Helmuth et al., PTRF '20)
- Méthodes variationnelles (Koehler et al., PMLR '22)
- Réseaux de neurones (Morningstar et Melko, JMLR '18)

Élaboration d'un nouveau projet

Alternatives aux MCMC pour la simulation de modèles sur réseau

Mesure de Gibbs : $\mu_\beta(\sigma) \propto \exp(-\beta \mathcal{H}(\sigma))$

avec $\mathcal{H}(\sigma) = \sum_{\{i,j\} \in \mathcal{E}} \sigma_i \sigma_j - h \sum_{i \in \mathcal{V}} \sigma_i$

Simulation : MCMC vs nouvelles techniques

- Pirogov-Sinai algorithmique (Helmuth et al., PTRF '20)
- Méthodes variationnelles (Koehler et al., PMLR '22)
- Réseaux de neurones (Morningstar et Melko, JMLR '18)

$$\mathcal{H}(\sigma) = J \sum_{\{i,j\} \in \mathcal{E}} (1 - \delta_{\sigma_i, \sigma_j}) + B \sum_{q \in \mathcal{Q}} \frac{(A_q - A_q^\star)^2}{2A_q^\star}$$

Merci de votre attention !

Merci de votre attention !