# Note SINDy

SINDy, ou "Sparse Identification of Nonlinear Dynamics", est une méthode d'identification automatique de modèle à partir de données, reposant sur un problème de minimisation construit pour assurer autant la fidélité aux données du modèle obtenu ainsi que sa "sparsité" (modèle parcimonieux).

Le modèle original de Brunton et al. (2015) a connu plusieurs variantes : PDE-FIND, SINDy with control, SINDy with sparse relaxed regularized regression et **SINDy-PI** (implicit-SINDy), que nous détaillons plus loin.

## 1 Modélisation fondamentale

On suppose que l'on dispose de données temporelles pour une fonction $x(t)$ sous la forme de deux vecteurs, $\boldsymbol{x}$ et $\dot{\boldsymbol{x}}$ de taille $N$. On suppose aussi qu'il existe un $f$ tel que

$$\dot{x}(t) = f(x(t)), \tag{1}$$

que l'on va tenter de reconstruire à partir des données. On considère une bibliothèque composée de $K$ fonctions candidates

$$\Theta = (\ f_1\ f_2 \dots\ ),$$

et l'on supposera que $f$ s'écrit comme *combinason linéaire parcimonieuse* de fonctions de cette bibliothèque. On construit la matrice de taille $N \times K$ associée à la bibliothèque et aux données :

$$\boldsymbol{\Theta}(\boldsymbol{x}) = \begin{pmatrix} \theta_1(\boldsymbol{x_1}) & \theta_2(\boldsymbol{x_1}) \dots \\ \theta_1(\boldsymbol{x_2}) & \theta_2(\boldsymbol{x_2}) \dots \\ \vdots & \vdots \end{pmatrix},$$

et le problème revient maintenant à trouver un vecteur $\xi \in \mathbb{R}^K$ de norme $\|\cdot\|_0$ la plus petite possible et qui minimise $\|\dot{\boldsymbol{x}} - \boldsymbol{\Theta}(\boldsymbol{x})\xi\|$.

**Problème de minimisation**

Précisément, le problème initial s'écrit donc

$$\underset{\xi \in \mathbb{R}^K}{\operatorname{argmin}} \|\dot{\boldsymbol{x}} - \boldsymbol{\Theta}(\boldsymbol{x})\xi\|_2 + \lambda \|\xi\|_0,$$

où $\lambda \geq 0$ es un paramètre de régularisation. Ce problème est néanmoins np-dur, ainsi on pourra le relaxer en le LASSO

$$\underset{\xi \in \mathbb{R}^K}{\operatorname{argmin}} \|\dot{\boldsymbol{x}} - \boldsymbol{\Theta}(\boldsymbol{x})\xi\|_2 + \lambda \|\xi\|_1.$$

**Sequential Thresholded Least Squares**

Le LASSO peut devenir cher computationellement pour les larges datasets, et de plus n'est pas très adapté à la sélection de coefficients. Une alternative proposée, qui se trouve de plus être simple et robuste au bruit, est la méthode des *Sequential Thresholded Least Squares*, où l'on imposera la sparsité "à la main", en effectuant récursivement une régression des moindres carrés, c'est à dire en résolvant $\underset{\xi \in \mathbb{R}^K}{\operatorname{argmin}} \|\dot{\boldsymbol{x}} - \boldsymbol{\Theta}(\boldsymbol{x})\xi\|_2$, puis en éliminant de la libraire $\Theta$ les fonctions pour lesquelles le coefficient associé est plus faible qu'un certain cut-off $\lambda$. L'identification de ce paramètre peut s'effectuer par de la validation croisée.

**Dimension** Si $x$ n'est pas à valeurs dans $\mathbb{R}$ mais $\mathbb{R}^d$ avec $d$ grand, la méthode SINDy deviendra moins adaptée et il faudra recourir à des méthodes de réduction de dimension. Pour notre problème, il faudra voir si c'est un souci ou non : ça dépendra de comment l'on gère les interactions.

**EDO à paramètres, forcing, contrôle** Il est possible d'étendre SINDy de (1) à des équations à paramètres, dependantes du temps ou avec forçage simplement en ajoutant des fonctions dans la librairie.

**Une illustration**

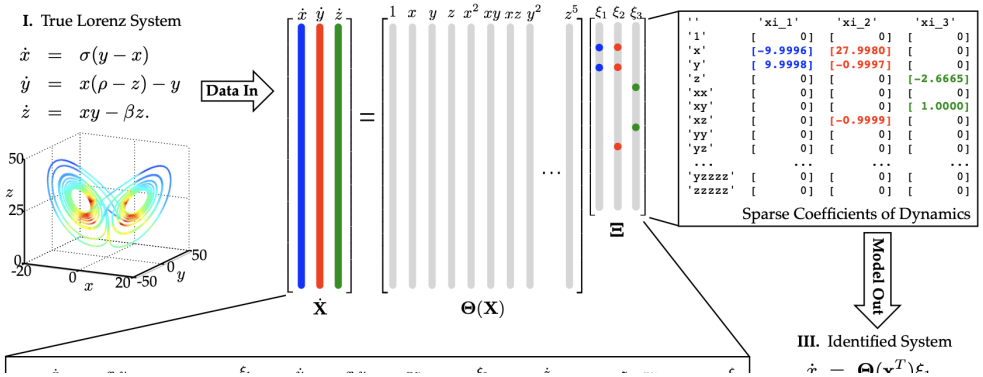


**Figure 1.** Représentation schématique de SINDy. *Brunton et al. 2015*

## 2 PDE-FIND

Une extension à SINDy pour retrouver des EDP a aussi été proposée dans cet article. L'extension est assez naturelle : supposons que l'on étudie un fonction $u(t,x)$, satisfaisant une EDP de la forme

$$\partial_t u = N(u, \partial_x u, \partial^2_{xx} u, \dots, x, t),$$

et que l'on cherche à retrouver la fonction $N$. La méthode PDE-FIND est une variante de SINDy où la bibliothèque $\Theta$ contiendra des fonctions de $u$, $\partial_x u$, $\partial^2_{xx} u$, $x$ et $t$.

## 3 SINDy with sparse regularized regression

La méthode SR3 proposée dans cet article propose une alternative à la méthode de Sequential Thresholded Least Squares (STLSQ). Elle en est une sorte de généralisation, bien que STLSQ ne soit pas réellement incluse.

Le problème de minimisation étudié sera ici

$$\underset{W\in\mathbb{R}^K}{\operatorname{argmin}} \quad \min_{\xi\in\mathbb{R}^K} \frac{1}{2}\|\dot{\boldsymbol{x}} - \boldsymbol{\Theta}(\boldsymbol{x})\xi\|_2^2 + \frac{1}{2\nu}\|\xi - W\|_2^2 + \lambda R(W)$$

Si l'on prend comme fonction $R$ la norme $\ell^0$, ce problème sera similaire à STLSQ. On peut commencer avec STLSQ, puis passer à SR3 si l'on est limité.

## 4 SINDy-PI, une méthode encore plus efficace

Le potentiel problème central de la méthode SINDy est que l'on doit supposer que la fonction $f$ est une *combinaison linéaire* d'éléments de la bibliothèque $\Theta$, ce qui restreint fortement les cas pour lesquels la méthode pourrait fonctionner.

L'idée proposée est alors de plutôt étudier le problème dit "implicite"

$$f(x, \dot{x}) = 0,$$

pour lequel on va alors tenter de reconstruire $f$ comme *combinaison linéaire parcimonieuse* de fonctions appartenant à la librairie $\Theta(x, \dot{x})$. À noter que le cas standard se retrouve avec la fonction $f(x, y) = y - g(x)$.

Dans ce cas, on considère alors une librairie $\Theta$ contenant des fonctions de $x$ et de $\dot{x}$, par exemple $\{1, x, x^2, \dot{x}, x\dot{x}, x^2\dot{x}\}$. Notons que cette librairie a alors accès aux équations comme $\dot{x} = \frac{x}{1+x}$.

**Implicit SINDy** Le problème, dans le cadre de SINDy, est alors de trouver le vecteur $\xi$ le plus sparse possible et tel que l'on ait

$$\boldsymbol{\Theta}(\boldsymbol{x}, \dot{\boldsymbol{x}})\, \xi \approx 0.$$

Ce problème est néanmoins mal posé, l'équation ci-dessus étant invariante par la multiplication d'un constante. Il faut donc fixer la valeur d'une norme de $\xi$ (par exemple sa norme 2). On obtient alors le problème de minimisation

$$\underset{\xi}{\text{argmin}} \quad \|\boldsymbol{\Theta}(\boldsymbol{x}, \dot{\boldsymbol{x}})\, \xi\|_2 + \lambda \|\xi\|_p, \qquad \|\xi\|_2 = 1,$$

où en théorie on souhaiterait $p = 0$ mais l'on mettrait en pratique $p = 1$. La résolution de ce problème se trouve être difficile, à cause de la contrainte, et l'on peut en fait appliquer une méthode plus efficace et robuste au bruit.

**SINDy-PI** L'accronyme "PI" vaut pour "Parallel, Implicit".

L'idée de SINDy-PI est de *traiter le problème implicite* décrit plus haut *avec la méthode explicite* traditionnelle. Considérons une librairie $\boldsymbol{\Theta}(\boldsymbol{x}, \dot{\boldsymbol{x}})$. Pour toute colonne $\boldsymbol{\theta}_j$ de $\boldsymbol{\Theta}$, on va résoudre le problème SINDy standard associé à l'équation

$$\boldsymbol{\theta}_j = \boldsymbol{\Theta}^j \xi^j,$$

où $\boldsymbol{\Theta}^j$ est la librairie $\boldsymbol{\Theta}$ à laquelle on a retiré la colonne $j$. Par suite, on fait la chose suivante : si le vecteur $\xi^j$ sélectionné n'est pas sparse et/ou donne une mauvaise prédiction, on rejette la fonction $\boldsymbol{\theta}_j$ et l'on recommence après avoir retiré $\boldsymbol{\theta}_j$ de la librairie. D'un autre côté, si l'on obtient un vecteur sparse et qui donne une bonne prédiction, on peut s'arrêter.

On peut aussi considérer deux librairies distinctes pour les termes de gauche et de droite.

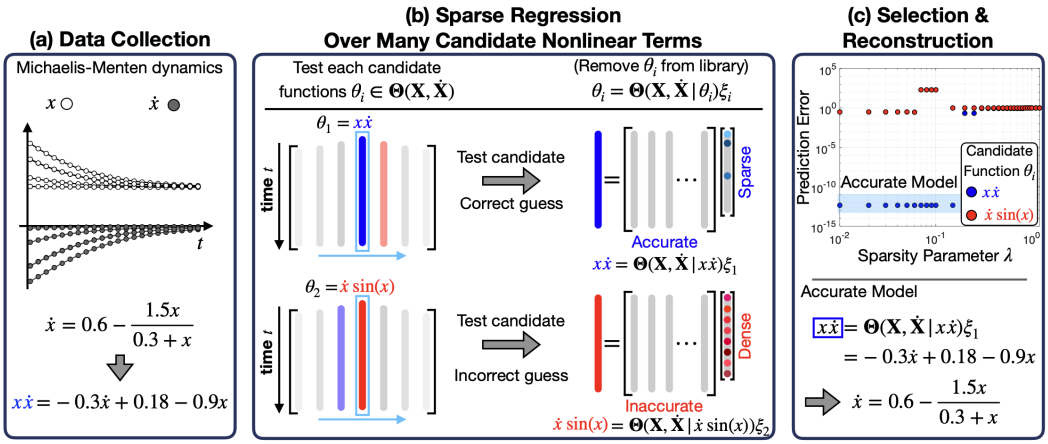


**Figure 2.** Représentation de SINDy-PI. *Kaheman et al. 2020*