Jackpot 2.0 : variété d'incertitude pour les problèmes inverses

Nathanaël MUNIER[1,2] Emmanuel SOUBIES[3,2] Pierre WEISS[2,3]

[1]IMT, Université de Toulouse, CNRS, France

[2]Centre de Biologie Intégrative, Laboratoire MCD, Université de Toulouse, CNRS, France

[3]IRIT, Université de Toulouse, CNRS, France

Résumé – Dans le contexte des problèmes inverses, nous avions introduit *Jackpot 1.0*, une méthode d'approximation de la région d'incertitude par une variété de dimension inférieure. Nous présentons une nouvelle méthode similaire qui détermine les points extrémaux des lignes de niveau de l'écart au modèle. Cette nouvelle méthode a deux avantages : elle repose uniquement sur un problème d'optimisation non contraint et elle renvoie des points qui maximisent le rapport entre écart de paramètre et écart de mesure. Cette méthode a été testée sur deux applications numériques de biologie des systèmes.

Abstract – In the context of inverse problems, we introduced *Jackpot 1.0*, a method for approximating the uncertainty region by a lower-dimensional manifold. We present a similar method which computes the extreme points of the level lines of the deviation from the model. This new method has two advantages: it is based solely on an unconstrained optimization problem and it returns points that maximize the ratio between parameter discrepancy and measurement discrepancy. This method has been tested on two numerical applications of systems biology.

1 Introduction

La résolution de problèmes inverses est au cœur de nombreux domaines scientifiques. Ces problèmes se caractérisent par la nécessité de reconstruire un ensemble de paramètres inconnus $x \in \mathbb{R}^N$ à partir d'un ensemble d'observations $y \in \mathbb{R}^M$, souvent incomplètes et bruitées. Mathématiquement, cette relation est modélisée par l'équation $y = \Phi(x) + b$ où $\Phi : \mathbb{R}^N \to \mathbb{R}^M$ représente le *modèle direct*, qui décrit comment les observations y sont générées à partir des paramètres x, et b correspond à un *bruit de mesure* ou à des incertitudes liées à l'acquisition des données. On suppose Φ différentiable dans ce papier.

Étant donné un estimateur $x^\star \in \mathbb{R}^N$ de x, obtenu à partir des observations y, nous nous intéressons à l'évaluation et à la quantification de l'incertitude associée à $x^\star$. Plus précisément, notre objectif est de développer des méthodes numériques permettant de caractériser la *région d'incertitude* définie pour $\varepsilon > 0$ par

$$\mathcal{U}^\varepsilon \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^N \;;\; \|\Phi(x) - \Phi(x^\star)\|_2^2 \leq \varepsilon \right\}. \qquad (1)$$

Cet ensemble procure un certain nombre d'informations sur le modèle. L'ensemble $\mathcal{U}^0$ permet d'analyser l'*identifiabilité structurelle* des paramètres $x^\star$. Si $\mathcal{U}^0$ ne contient que $x^\star$, alors le modèle est structurellement identifiable, c'est-à-dire qu'il ne présente aucune redondance au niveau des paramètres. Pour $\varepsilon > 0$, l'ensemble $\mathcal{U}^\varepsilon$ caractérise l'*identifiabilité pratique*. Son étendue et sa forme renseignent sur l'incertitude associée aux paramètres. Plus l'ensemble $\mathcal{U}^\varepsilon$ est grand, plus le modèle est mal posé. Pour la suite, on notera la fonction d'écart quadratique par

$$F(x) \stackrel{\text{def}}{=} \frac{1}{2}\|\Phi(x) - \Phi(x^\star)\|_2^2. \qquad (2)$$

Les méthodes actuelles. On peut regrouper les méthodes d'analyse d'incertitude en deux classes selon qu'elles sont déterministes ou probabilistes. La plupart des méthodes déterministes existantes sont dédiées à l'identifiabilité structurelle et ne sont applicables qu'à des modèles algébriques (e.g., DAISY [2], COMBOS [9], STRIKE-GOLDD [12]). D'autres méthodes approchent la région d'incertitude à l'aide d'intervalles pour chaque coordonnée (profil de vraisemblance) [11]. Pour une étude plus approfondie sur ces méthodes, nous renvoyons le lecteur vers l'article [13]. De leur côté, les approches probabilistes s'appuient sur des méthodes d'échantillonnage telles que les méthodes de bootstrap [3] ou de Monte-Carlo par chaînes de Markov [1]. Après échantillonnage, la région d'incertitude peut alors être visualisée et approchée par un ensemble plus simple comme des ellipsoïdes [4] ou des polyèdres [7].

Difficulté du problème. La difficulité du problème tient dans la forme de la région d'incertitude. C'est un ensemble défini implicitement, de dimension pleine N et qui n'est pas forcément connexe. À moins de l'approcher par un ensemble plus simple ou de dimension plus faible, il est difficile à représenter ou à visualiser. Les méthodes probabilistes sont confrontées à la difficulté de l'échantillonnage qui souffre du fléau de la dimension. Les méthodes déterministes algébriques, quant à elles, sont limitées au cadre algébrique et peuvent être incapables de rendre la complexité de la région d'incertitude qu'elles cherchent à paramétrer. Ces méthodes sont donc généralement limitées à des problèmes en petite dimension. Pour les méthodes reposant sur le calcul de profils de vraisemblance, un intervalle doit être déterminé pour chaque paramètre, limitant également leur application à des problèmes en faible dimension. De plus, ces méthodes ne permettent pas de coupler les différents paramètres.

Toutes ces difficultés nous conduisent à concevoir de nouvelles méthodes d'approximation de $\mathcal{U}^\varepsilon$ répondant aux objectifs suivants.

Objectifs. Déterminer une variété $\mathcal{M}$ de faible dimension $D \ll N$ qui approche $\mathcal{U}^\varepsilon$ et vérifie les propriétés suivantes :

- Déterministe : la variété obtenue dépend de manière déterministe de la fonction F. En particulier, cette variété n'a recours à aucun échantillonnage pour être construite.
- Locale : la variété est construite en n'utilisant que des informations géométriques locales de F.

Nous avons récemment proposé une méthode numérique, nommée *Jackpot 1.0* (Jacobian Kernel Projection Optimization), permettant de déterminer une telle variété de faible dimension [10]. Nous la décrivons brièvement dans la section 2 et évoquons ses limitations. Nous présentons ensuite (section 3) une variante permettant de dépasser certaines limitations.

2 La variété Jackpot 1.0

Description. Soit $A \in \mathbb{R}^{N\times D}$ la matrice formée par les D vecteurs singuliers de la jacobienne de Φ en $x^\star$, associés aux valeurs singulières les plus faibles. La variété Jackpot 1.0 est paramétrée, pour $u \in \mathbb{R}^D$ par

$$\phi_A(u) \stackrel{\text{def}}{=} \operatorname*{argmin}_{x\in\mathcal{L}_u} F(x) \tag{3}$$

où $\mathcal{L}_u$ est un espace affine de dimension $(N-D)$ défini par

$$\mathcal{L}_u = x^\star + Au + \mathrm{Im}(A)^\perp. \tag{4}$$

Une illustration de cette méthode est donnée pour les cas $D = 1$ et $D = 2$ sur la figure 1 (a) et (b). La variété Jackpot 1.0 est alors

$$\mathcal{M}_1^\varepsilon \stackrel{\text{def}}{=} \left\{ \phi_A(u)\ ;\ u \in \mathbb{R}^D, F(\phi_A(u)) < \frac{1}{2}\varepsilon^2 \right\}. \tag{5}$$

Limitations. Lorsque la dimension N augmente, la matrice jacobienne de Φ n'est plus stockable en mémoire. Des méthodes « matrix-free » sont alors nécessaires mais leur convergence vers les vecteurs singuliers peut être lente, en particulier dans le cas d'une jacobienne mal conditionnée. La variété Jackpot dépend de cette matrice A et peut présenter une instabilité. Par exemple, la figure 2 (colonne de gauche) compare pour $D = 1$ les variétés obtenues avec Jackpot lorsque A est construite à partir d'une perturbation du vecteur propre associé à la plus petite valeur singulière de Jac_Φ. On observe que les différentes trajectoires sont différentes. Ainsi, les difficultés de calcul des vecteurs singuliers mentionnées précédemment (en pratique pour des problèmes en grande dimension, seule une approximation de ces derniers est possible) impactent la variété obtenue.

Un des objectifs de la prochaine section est de modifier la méthode Jackpot afin de limiter sa sensibilité à la matrice A.

3 La variété Jackpot 2.0

L'idée de prendre l'espace linéaire $\mathrm{Im}(A)$ (avec A comme définie dans la section précédente) caractérisant les plus faibles variations de F au voisinage de $x^\star$ est naturelle. Néanmoins, la projection selon l'orthogonal n'est pas la seule manière de faire et d'autres formes peuvent être envisageables. Dans cette section, nous proposons de prendre une contrainte sphérique et nous montrerons que c'est celle qui fournit les points les plus extrémaux pour une ligne de niveau de F fixée.

3.1 Définition

Soit $A \in \mathbb{R}^{N\times D}$ la matrice formée par les D vecteurs singuliers de la jacobienne de Φ en $x^\star$, associés aux valeurs singulières les plus faibles. Nous proposons la paramétrisation suivante

$$\psi_A(u) \stackrel{\text{def}}{=} \operatorname*{argmin}_{x\in\mathcal{C}_u} F(x) \tag{6}$$

où $\mathcal{C}_u$ est une sphère de dimension $(N-D)$ définie par

$$x \in \mathcal{C}_u \iff \|x - x^\star\|_2^2 = \|u\|_2^2 \ \text{ et } \ x \in \mathcal{P}_u \tag{7}$$

avec $\mathcal{P}_u$ le sous-espace défini par

$$\mathcal{P}_u = x^\star + \mathrm{Vect}(Au) \oplus \mathrm{Im}(A)^\perp. \tag{8}$$

La variété Jackpot 2.0 est alors

$$\mathcal{M}_2^\varepsilon \stackrel{\text{def}}{=} \left\{ \psi_A(u)\ ;\ u \in \mathbb{R}^D, F(\psi_A(u)) < \frac{1}{2}\varepsilon^2 \right\}. \tag{9}$$

Une illustration de cette nouvelle méthode est donnée pour les cas $D = 1$ et $D = 2$ sur la figure 1 (c) et (d). Cette définition est motivée par la remarque suivante :

Remarque 3.1. *Dans le cas $D = 1$, la variété $\mathcal{M}^\varepsilon$ est indépendante du choix de $A \in \mathbb{R}^{N\times D}$, car $\mathcal{P}_u = \mathbb{R}^N$.*

Ce résultat est illustré sur la figure 2 (colonne de droite).

3.2 Calcul numérique

L'ensemble $\mathcal{C}_u$ étant simple, une paramétrisation possible du problème (6) sous forme non contrainte est :

$$\eta(u) \stackrel{\text{def}}{=} \operatorname*{argmin}_{x\in\mathbb{R}^N} F(x^\star + \alpha Au + \Pi_\perp x) \tag{10}$$

où $\alpha^2 = 1 - \|\Pi_\perp x\|_2^2 / \|u\|_2^2$ et $\Pi_\perp \stackrel{\text{def}}{=} \mathrm{Id}_N - AA^T$. On obtient alors :

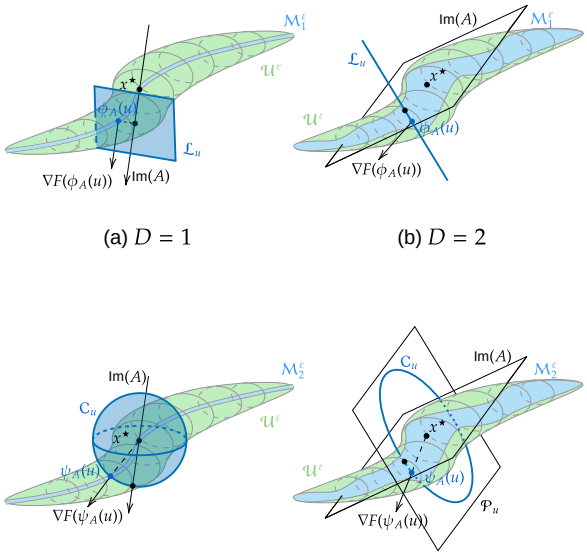
$$\psi_A(u) = x^\star + \alpha Au + \Pi_\perp \eta(u). \tag{11}$$

L'algorithme permettant de calculer la variété proposée est donnée par le pseudo-code de l'algorithme 1. En dimension $D = 1$, la matrice A peut être quelconque contrairement à l'algorithme Jackpot 1.0. Pour résoudre le problème d'optimisation à la ligne 5, le problème étant non contraint, une méthode d'ordre 1 ou de quasi-Newton peut être appliquée.

3.3 Propriétés de la variante Jackpot 2.0

Une première propriété assez simple à établir est l'inclusion des variétés lorsqu'on augmente la dimension D :

Proposition 3.1. *Soit $\mathcal{M}^{(D)}$ la variété de dimension D obtenue par un des algorithmes Jackpot, en ajoutant une colonne à la matrice A. Dans ce cas : $\mathcal{M}^{(1)} \subset \mathcal{M}^{(2)} \subset \mathcal{M}^{(3)} \subset \ldots$.*



(a) $D = 1$ (b) $D = 2$

(c) $D = 1$ (d) $D = 2$

FIGURE 1 : Illustration de la variété Jackpot 1.0 (a et b) et Jackpot 2.0 (c et d) approchant la région d'incertitude $\mathcal{U}^\varepsilon$. Sur l'ensemble $\mathcal{L}_u$ (droite ou plan) ou sur l'ensemble $\mathcal{C}_u$ (cercle ou sphère), le point de la variété $\mathcal{M}$ est donnée par le minimiseur de F. En ce point, le gradient de F appartient à $\mathrm{Im}(A)$ (pour Jackpot 1.0) ou bien son projeté sur $\mathcal{P}_u$ pointe vers $\psi_A(u)$ en partant de $x^\star$ (pour Jackpot 2.0). On notera que dans le cas où $D = 1$, l'espace $\mathcal{P}_u$ est $\mathbb{R}^N$ entier.

Algorithme 1 : Exploration de la variété

Entrées : $x^\star \in \mathbb{R}^N$, $1 \leq D < N$, une grille $\mathcal{G}$ de $\mathbb{R}^D$
Choisir $A \in \mathbb{R}^{N\times D}$ de colonnes orthonormales
(e.g. les D vect. sing. les plus à droite de Jac_Φ)
pour u_k *ordonné par Breadth-First-Search* **faire**
    Calculer $\eta(u_k)$ en partant de $u_{k-1}(10)$.
    En déduire $\psi_A(u_k)$ de (11).
fin
Sortie : la variété discrétisée $(\psi_A(u_k))_{u\in\mathcal{G}}$.

Il est instructif de reformuler le problème (6) où le rôle de contrainte et de fonction coût sont inversées :

$$\tilde{\psi}_A(u) \stackrel{\text{def}}{=} \underset{F(x)=\|u\|_2^2,\ x\in\mathcal{P}_u}{\operatorname{argmax}} \|x - x^\star\|_2 . \tag{12}$$

On sait déjà que ces deux problèmes (12) et (6) partagent les mêmes points critiques. Cette reformulation permet en particulier d'observer la propriété suivante :

Proposition 3.2. *Supposons que les problèmes* (12) *et* (6) *partagent, en plus de leurs points critiques, les mêmes minima/maxima locaux. Posons l'ensemble des points extrémaux*

$$\Gamma = \{\gamma(r), r \geq 0\} \quad \textit{avec} \quad \gamma(r) = \underset{F(x)=r}{\operatorname{argmax}} \|x - x^\star\|_2. \tag{13}$$

Alors en toute dimension D, la variété Jackpot 2.0 contient Γ.[1]

[1] D'un point de vue numérique, ce résultat suppose qu'on est capable de calculer exactement le maximum global (12), ce qui n'est vrai que localement autour de $x^\star$.

Finalement, nous pouvons lier Jackpot 1.0 à Jackpot 2.0 de la façon suivante.

Proposition 3.3. *Si le modèle Φ est non identifiable sur une variété $\mathcal{M}_\Phi$ de dimension D, telle que :*

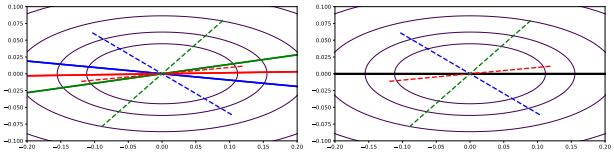
$$\forall x \in \mathcal{M}_\Phi, \qquad F(x) = 0$$
$$\forall x \notin \mathcal{M}_\Phi, \qquad F(x) > 0$$

alors les variétés obtenues par Jackpot 1.0 et 2.0 coïncident avec $\mathcal{M}_\Phi$.

4 Expériences numériques

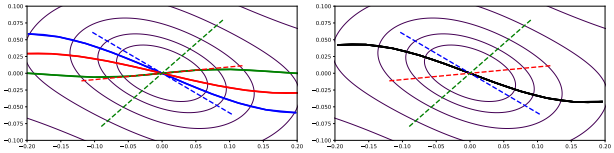
4.1 Comparaison des méthodes

$$\Phi(x) = \Delta x \stackrel{\text{def}}{=} \mathrm{Diag}(0.4, 1)x$$



(a) Jackpot 1.0 (b) Jackpot 2.0

$$\Phi(x) = \Delta x + 10^{-1}\sin(2\pi R x)$$



(c) Jackpot 1.0 (d) Jackpot 2.0

FIGURE 2 : Comparaison de Jackpot 1.0 (à gauche) et Jackpot 2.0 (à droite) pour différentes valeurs de A. Ici $D = 1$ et $N = 2$. Les lignes de niveau de Φ sont représentées en foncé. Les lignes discontinues colorées représentent les différentes directions A. Les lignes pleines colorées représentent les variétés $\mathcal{M}$ obtenues pour différentes directions. La courbe Jackpot 2.0 est identique pour chaque direction A et est représentée en noir. Dans le deuxième exemple, R est la matrice de rotation d'angle $\frac{\pi}{2}$.

En dimension 1, l'intérêt de la version 2.0 de Jackpot par rapport à la version 1.0 [10] est clair : la variante proposée est indépendante de la direction A choisie. Cela est illustré sur des exemples en dimension 2 sur la figure 2. Il n'y a donc plus nécessité de calculer exactement les couples de valeurs-vecteurs singuliers les plus faibles de la jacobienne de Φ au point initial $x^\star$. Cette propriété témoigne d'une stabilité accrue de la méthode Jackpot 2.0.

4.2 Des modèles dynamiques

Le modèle SIR Le modèle SIR (Sains-Infectés-Rétablis) [6] décrit l'évolution au cours du temps de la proportion infectée d'une population lors d'une épidémie. D'abord pris comme outil théorique, il a été utilisé numériquement pour la première fois dans les années 1980 afin de modéliser l'épidémie du

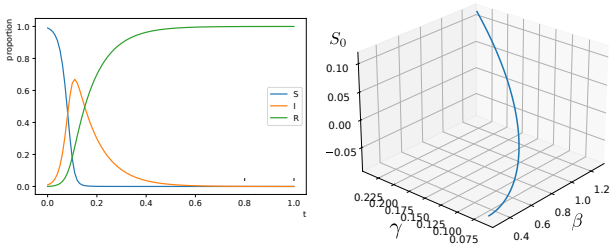
SIDA. Les équations qui le décrivent sont données par

$$\begin{cases} \dot{S} = & -\beta S \cdot I, \\ \dot{I} = & \beta S \cdot I - \gamma I, \\ \dot{R} = & \gamma I, \end{cases} \tag{14}$$

où les paramètres sont $x = (S_0, \beta, \gamma)$. Pour l'application numérique de ce papier nous prendrons les valeurs $x = (0.01, 0.8, 0.1)$ La sortie du modèle est l'évolution de la proportion infectée I à intervalle de temps régulier comme décrit sur la figure 3 (a). On peut remarquer que R n'est pas considéré ici. Une discrétisation d'Euler explicite de $K = 100$ points avec un pas de temps de $dt = 1$ a été utilisée.

Proposition 4.1. *Le système du SIR est non-identifiable et $\mathcal{U}^0$ est la courbe décrite par $\beta S_0 - \gamma$ constant.*

Comme montré figure 3 (b), l'algorithme Jackpot 2.0 retrouve numériquement cette non-identifiabilité. Le choix $D = 1$ vient du fait qu'il s'agit théoriquement d'une courbe.



(a) Courbe de l'évolution des proportions de la population

(b) Courbe Jackpot 2.0

FIGURE 3 : Le modèle du SIR. La courbe (b) générée par l'algorithme Jackpot 2.0 coïncide exactement avec la courbe $\beta S_0 - \gamma$ constant.

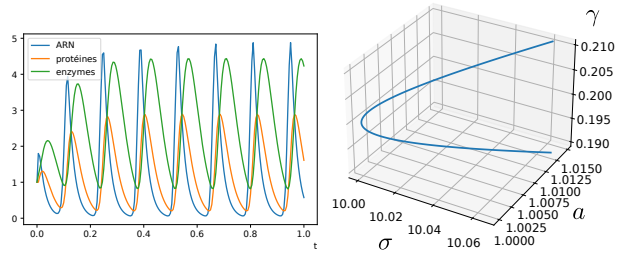
Le modèle de Goodwin Le modèle de Goodwin [5] décrit le mécanisme de l'expression périodique des protéines par une boucle de rétroaction négative qui inhibe la transcription de l'ARN messager. En 2005, l'équipe de Locke [8] a montré que l'horloge circadienne peut être modélisée par le modèle de Goodwin. L'équation aux dérivées ordinaires est décrite ci-dessous :

$$\begin{cases} \dot{y} = & \frac{a}{c+w^\sigma} - by, \\ \dot{z} = & \alpha y - \beta z, \\ \dot{w} = & \gamma z - \delta w. \end{cases} \tag{15}$$

Les paramètres de ce modèle sont $x = (a, b, c, \sigma, \alpha, \beta, \gamma, \delta)$ et pour cet exemple, nous prendrons les valeurs $x = (1, 0.1, 1, 10, 0.1, 0.1, 0.2, 0.1)$ et $(y_0, z_0, w_0) = (1, 1, 1)$. Une discrétisation d'Euler explicite de $K = 200$ points avec un pas de temps de $dt = 2$ a été utilisée. Ce modèle n'est pas identifiable [12] si on observe uniquement y, mais la structure de non identifiabilité n'a pas d'expression analytique. L'algorithme retrouve cette non-identifiabilité. Un écart de 10^{10} entre la dernière et les autres valeurs singulières de la jacobienne indique une bonne approximation de $\mathcal{U}^\varepsilon$ par une variété de dimension $D = 1$ comme affichée à la figure 4.

5 Conclusion

Dans ce travail, nous avons présenté des méthodes permettant d'approcher numériquement la région d'incertitude $\mathcal{U}^\varepsilon$ avec



(a) Courbe de l'évolution des différentes entités

(b) Courbe Jackpot 2.0

FIGURE 4 : Le modèle de Goodwin. En (b) est représenté la courbe Jackpot 2.0 selon les trois coordonnées (a, σ, γ) où la variance est la plus forte.

une variété de faible dimension. Nous avons notamment proposé une variante de la méthode Jackpot [10] que nous avons introduite récemment, et mis en avant une propriété de stabilité importante. En particulier, la variété Jackpot 2.0 contient systématiquement un ensemble Γ géométrique qui maximise l'erreur d'entrée à erreur de sortie fixée.

Remerciements Ce travail a bénéficié d'une aide de l'École Universitaire de Recherche MINT (ANR-18-EURE- 0023), de l'ANR Micro-Blind ANR-21- CE48-0008 et des ressources HPC de GENCI-IDRIS (Grant 2021-AD011012210R3).

Références

[1] J. M. BARDSLEY : MCMC-based image reconstruction with uncertainty quantification. *SIAM Journal on Scientific Computing*, 2012.

[2] G. BELLU, M. P. SACCOMANI, S. AUDOLY et L. D'ANGIÒ : DAISY : A new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, 2007.

[3] T. ENDO, T. WATANABE et A. YAMAMOTO : Confidence interval estimation by bootstrap method for uncertainty quantification using random sampling method. *Journal of Nuclear Science and Technology*, 2015.

[4] F. GOLESTANEH, P. PINSON, R. AZIZIPANAH-ABARGHOOEE et H. B. GOOI : Ellipsoidal prediction regions for multivariate uncertainty characterization. *IEEE Transactions on Power Systems*, 2018.

[5] B. C. GOODWIN : Oscillatory behavior in enzymatic control processes. *Advances in enzyme regulation*, 1965.

[6] T. HARKO, F. SN. LOBO et M. K. MAK : Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation*, 2014.

[7] R. J. HYNDMAN : Computing and graphing highest density regions. *The American Statistician*, 1996.

[8] J. CW. LOCKE, A. J. MILLAR et M. S. TURNER : Modelling genetic networks with noisy and varied experimental data : the circadian clock in arabidopsis thaliana. *Journal of theoretical biology*, 2005.

[9] N. MESHKAT, C. E. KUO et J. DISTEFANO III : On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and combos : a novel web implementation. *PloS one*, 2014.

[10] N. MUNIER, E. SOUBIES et P. WEISS : Jackpot : Approximating uncertainty domains with adversarial manifolds. 2024.

[11] S. A. MURPHY et A. W. Van der VAART : On profile likelihood. *Journal of the American Statistical Association*, 2000.

[12] A. F. VILLAVERDE, A. BARREIRO et A. PAPACHRISTODOULOU : Structural identifiability of dynamic systems biology models. *PLoS computational biology*, 2016.

[13] F.-G. WIELAND, A. L. HAUBER, M. ROSENBLATT, C. TÖNSING et J. TIMMER : On structural and practical identifiability. *Current Opinion in Systems Biology*, 2021.