Le principe de longueur de description minimale pour la modélisation des données

ou la théorie statistique de l'information pour bien exploiter les mesures

par **François CHAPEAU-BLONDEAU**
Faculté des Sciences
Université d'Angers - 49000 Angers
chapeau@univ-angers.fr

RÉSUMÉ

Le principe de longueur de description minimale pour la modélisation des données à partir de mesures bruitées est exposé. Cette méthode, basée sur un critère informationnel, permet, pour une classe quelconque de modèles spécifiée à nombre variable de paramètres, d'estimer à la fois le nombre optimal de paramètres ainsi que leurs valeurs optimales. Le principe est illustré à l'aide de plusieurs exemples de régression polynomiale, mis en œuvre avec le logiciel libre de calcul scientifique Scilab, où sont déterminés à la fois le degré optimal du polynôme ainsi que les valeurs optimales de ses coefficients.

1 Introduction

Une grandeur physique y est supposée dépendre d'une autre grandeur physique x, selon un modèle que l'on va chercher à estimer. On dispose d'un ensemble $\mathcal{M}$ de N mesures (x_n, y_n), $n = 1$ à N, i.e.

$$\mathcal{M} = \{(x_n, y_n),\ n = 1, \dots N\} \ . \tag{1}$$

Un y_n donné est sous la dépendance du x_n associé, selon une loi fonctionnelle sous-jacente notée $Q_K(.)$, à laquelle se superpose une perturbation ou bruit de mesure, conformément à la relation

$$y_n = Q_K(x_n) + \xi_n \ , \tag{2}$$

Dans l'Éq. (2), la quantité ξ_n représente un échantillon de bruit, ou plus généralement tout écart existant entre la valeur théorique modélisée $Q_K(x_n)$ et l'observation y_n.

La loi ou modèle $Q_K(.)$ est paramétré par K paramètres a_k, pour $k = 1$ à K ; on appelle K l'ordre du modèle. Il peut s'agir par exemple d'un modèle polynomial réalisant

$$Q_K(x) = \sum_{k=1}^{K} a_k x^{k-1} \ , \tag{3}$$

ou bien $Q_K(.)$ peut être une fraction rationnelle (rappport de deux polynômes) à K paramètres, ou bien une superposition de sinusoïdes chacune à 3 paramètres (amplitude, fréquence, phase

à l'origine), ou bien une superposition de gaussiennes chacune à 3 paramètres (position, amplitude, largeur), ou bien tout autre modèle paramétrique concevable. On fixe ainsi une forme spécifiée pour un modèle à K paramètres.

À partir de l'ensemble $\mathcal{M}$ des N mesures (x_n, y_n) de l'Éq. (1), deux questions peuvent alors être abordées :
(1) On suppose connu l'ordre K du modèle. On souhaite utiliser l'ensemble $\mathcal{M}$ des mesures afin d'estimer "au mieux" les K paramètres a_k du modèle.
(2) On ne connaît pas l'ordre K du modèle. On souhaite utiliser l'ensemble $\mathcal{M}$ des mesures afin d'estimer d'abord l'ordre K du modèle qui modélise "au mieux" les mesures, puis les K paramètres a_k du modèle.

2 Estimation des paramètres à ordre du modèle connu

On suppose donc ici que l'ordre K du modèle est connu. On souhaite utiliser l'ensemble $\mathcal{M}$ des mesures afin d'estimer "au mieux" les K paramètres a_k du modèle. On note $\boldsymbol{a} = (a_1, \dots a_K)$ le vecteur des K paramètres inconnus à estimer.

Pour un x_n donné, l'échantillon de bruit ξ_n est modélisé comme une variable aléatoire, et alors le y_n résultant via l'Éq. (2) est vu aussi comme une variable aléatoire. On introduit $p(\mathcal{M}|\boldsymbol{a})$, la probabilité d'observer l'ensemble $\mathcal{M}$ des mesures sachant que le vecteur des paramètres vaut $\boldsymbol{a}$.

Une procédure classique existe pour estimer les paramètres $\boldsymbol{a}$, qui constitue a priori un choix raisonnable, et qui possède des propriétés théoriques remarquables. Il s'agit de la procédure du maximum de vraisemblance [1, 2, 3]. Lorsque l'on est en possession d'un jeu de mesures $\mathcal{M}$, on estime les paramètres $\boldsymbol{a}$ qui l'ont produit, en choisissant pour $\boldsymbol{a}$ la valeur qui maximise la probabilité $p(\mathcal{M}|\boldsymbol{a})$. Si l'on note $\widehat{\boldsymbol{a}}_{MV}$ les valeurs des paramètres ainsi estimées, on a donc

$$\widehat{\boldsymbol{a}}_{\mathrm{MV}} = \arg\max_{\boldsymbol{a}} \; p(\mathcal{M}|\boldsymbol{a}) \; . \tag{4}$$

L'estimation du maximum de vraisemblance choisit donc, pour les paramètres, les valeurs qui maximisent la probabilité d'observer les mesures que l'on a en main.

Une autre interprétation, qui nous sera utile pour la suite, de l'estimation par maximum de vraisemblance, est offerte par la théorie statistique de l'information [4, 5]. Pour un événement de probabilité $p(\mathcal{M}|\boldsymbol{a})$ (la probabilité d'observer nos mesures pour un vecteur de paramètres $\boldsymbol{a}$ donné), on définit $-\log p(\mathcal{M}|\boldsymbol{a})$ comme l'information associé à cet événement. Cette définition qui formalise une notion d'information de façon probabiliste, possède des conséquences très profondes. En particulier, on montre que l'information $-\log p(\mathcal{M}|\boldsymbol{a})$ mesure aussi la longueur du codage le plus court qui peut être réalisé pour cet événement. Choisir les paramètres pour maximiser la probabilité $p(\mathcal{M}|\boldsymbol{a})$ des mesures, comme dans l'Éq. (4), revient à minimiser l'information $-\log p(\mathcal{M}|\boldsymbol{a})$ apportée par les mesures une fois les paramètres connus, ou encore à minimiser la longueur du codage des mesures $L(\mathcal{M}|\boldsymbol{a}) = -\log p(\mathcal{M}|\boldsymbol{a})$ une fois les paramètres connus. L'estimation des paramètres du modèle au sens du maximum de vraisemblance selon l'Éq. (4), est donc équivalente à l'estimation au sens de la longueur de description (codage) minimale, exprimée par

$$\widehat{\boldsymbol{a}}_{\mathrm{MV}} = \arg\min_{\boldsymbol{a}} \; L(\mathcal{M}|\boldsymbol{a}) = \arg\min_{\boldsymbol{a}} \; -\log p(\mathcal{M}|\boldsymbol{a}) \; . \tag{5}$$

Cette longueur de description minimale est ce que l'on escompte d'un bon paramétrage du modèle : il décrit bien en général les mesures, et une fois qu'on le possède, il n'y a pas beaucoup d'information à ajouter pour décrire les particularités d'un jeu de mesures donné.

Pour la forme de l'Éq. (2), c'est la spécification des probabilités des ξ_n, lesquels nous appelons ici échantillons de bruit, qui confère sa structure probabiliste au problème et qui détermine la probabilité $p(\mathcal{M}|\boldsymbol{a})$. Pour aller plus loin et être illustratif, nous considérons une situation fréquemment rencontrée en pratique, où les échantillons de bruit ξ_n, pour des valeurs de n distinctes, sont des variables aléatoires indépendantes et identiquement distribuées selon la densité de probabilité $f_\xi(u)$. En vertu de l'Éq. (2), il vient alors

$$p(\mathcal{M}|\boldsymbol{a}) = \prod_{n=1}^{N} f_\xi\Big(y_n - Q_K(x_n)\Big) \,, \tag{6}$$

soit

$$L(\mathcal{M}|\boldsymbol{a}) = -\log p(\mathcal{M}|\boldsymbol{a}) = -\sum_{n=1}^{N} \log f_\xi\Big(y_n - Q_K(x_n)\Big) \,. \tag{7}$$

En spécifiant $f_\xi(u)$, et une fois en possession d'un jeu de N mesures (x_n, y_n), l'Éq. (7) représente une fonction des K paramètres $(a_1, \ldots a_K) = \boldsymbol{a}$, pour laquelle il convient de rechercher, conformément au principe de l'Éq. (5), le minimiseur $\widehat{\boldsymbol{a}}_{MV}$ qui la minimise. On peut pour cela utiliser toute méthode de minimisation d'une fonction à plusieurs variables [6].

Une situation très fréquemment rencontrée en pratique est le cas du bruit gaussien, de moyenne nulle et d'écart-type σ, associé à la densité de probabilité

$$f_\xi(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{u^2}{2\sigma^2}\right) \,. \tag{8}$$

Dans ce cas l'Éq. (7) donne

$$-\log p(\mathcal{M}|\boldsymbol{a}) = N\log(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(y_n - Q_K(x_n)\Big)^2 \,. \tag{9}$$

L'estimation des paramètres du modèle selon le principe de longueur de description minimale de l'Éq. (5) est alors équivalente à

$$\widehat{\boldsymbol{a}}_{MV} = \arg\min_{\boldsymbol{a}} \; -\log p(\mathcal{M}|\boldsymbol{a}) \;=\; \arg\min_{\boldsymbol{a}} \; \sum_{n=1}^{N}\Big(y_n - Q_K(x_n)\Big)^2 \,. \tag{10}$$

L'équation (10) montre que, dans le cas du bruit gaussien, l'estimation au sens de la longueur de description minimale (ou équivalemment au sens du maximum de vraisemblance) est équivalente à l'estimation au sens des moindres carrés.

L'estimation au sens des moindres carrés [3] détermine les paramètres $\boldsymbol{a}$ du modèle en minimisant l'écart quadratique entre les valeurs mesurées (les y_n) et les valeurs prédites par le modèle (les $Q_K(x_n)$). Il s'agit aussi d'une approche générale pour l'estimation, qui ne s'appuie pas sur une formalisation probabiliste de la nature des écarts entre les valeurs mesurées (les y_n) et les valeurs prédites par le modèle déterministe (les $Q_K(x_n)$). Elle s'appuie néanmoins sur une mesure d'écart, quelque peu ad hoc ou arbitraire, l'écart quadratique $\sum_{n=1}^{N}[y_n - Q_K(x_n)]^2$. D'autres mesures d'écart pourraient a priori être également valables, comme $\sum_{n=1}^{N}|y_n - Q_K(x_n)|$ ou $\sum_{n=1}^{N}|y_n - Q_K(x_n)|^\alpha$ pour tout $\alpha > 0$. Minimiser l'une ou l'autre de ces mesures d'écart conduit en général à des valeurs différentes des paramètres $\boldsymbol{a}$ estimés.

Dans le cas de l'estimation par longueur de description minimale, la fonction à minimiser découle toujours d'un même principe (probabiliste) général. Ce principe conduit, dans le cas

du bruit gaussien, comme on l'a vu, à minimiser l'écart quadratique. Avec un bruit Laplacien de densité $f_\xi(u)$ en $\exp(-\beta|u|)$, le principe de longueur de description minimale conduit à minimiser, non plus l'écart quadratique, mais $\sum_{n=1}^{N}|y_n - Q_K(x_n)|$; avec un bruit gaussien généralisé de densité $f_\xi(u)$ en $\exp(-\beta|u|^\alpha)$, il conduit à minimiser $\sum_{n=1}^{N}|y_n - Q_K(x_n)|^\alpha$.

Clairement, on a ici affaire à deux philosophies distinctes pour l'estimation paramétrique, où la conceptualisation ne réside pas au même endroit. Soit on conceptualise pour justifier un critère d'écart à minimiser, soit on conceptualise pour établir une formalisation probabiliste d'où découlera mécaniquement le critère à minimiser (une fois admis le principe de longueur de description minimale). Le principe de longueur de description minimale offre néanmoins quelquechose de plus : il peut aussi s'appliquer au problème de l'estimation à nombre de paramètres inconnus [7, 8].

3 Estimation de l'ordre du modèle et des paramètres

On suppose maintenant que l'ordre K du modèle est inconnu. On souhaite utiliser l'ensemble $\mathcal{M}$ des mesures afin d'estimer d'abord l'ordre K du modèle qui modélise "au mieux" les mesures, puis les K paramètres a_k du modèle.

Selon un principe d'économie, appelé aussi parfois rasoir d'Occam, un bon modèle doit représenter adéquatement les données sans toutefois multiplier de façon excessive les paramètres. Par exemple, dans le cas de la régression polynomiale de l'Éq. (3), si l'on possède N points de mesure (x_n, y_n), on sait qu'en général avec un nombre $K = N$ de coefficients polynomiaux a_k, on peut modéliser de façon exacte les mesures, en réduisant à zéro par exemple toute mesure d'écart ou d'erreur du type $\sum_{n=1}^{N}|y_n - Q_K(x_n)|^\alpha$. Cependant, pour N mesures bruitées, un tel modèle à N paramètres constitue bien souvent un modèle trop compliqué, qui s'épuise à représenter les accidents d'une réalisation particulière du bruit, et manque les régularités reproductibles d'une loi "simple" sous-jacente. L'apport de points de mesure ne conduit qu'à l'ajout de paramètres au modèle, au lieu de progressivement consolider un modèle simple sous-jacent. Ainsi, autant de paramètres que de points de mesure ne constitue pas en général une forme satisfaisante pour le modèle, bien que ceci puisse être associé à une erreur de modélisation nulle. Un modèle satisfaisant présentera une forme de "simplicité" en équilibrant deux exigences : une erreur de modélisation suffisamment faible associée à un nombre économe de paramètres. Le principe de longueur de description minimale offre une formalisation rigoureuse de ces considérations [7, 8].

Comme on l'a vu dans la section précédente, une fois les K paramètres a_k connus, la longueur du code pour décrire un jeu de mesures $\mathcal{M}$ est $L(\mathcal{M}|\boldsymbol{a}) = -\log p(\mathcal{M}|\boldsymbol{a})$. On doit maintenant compléter avec la longueur du code $L(\boldsymbol{a})$ destiné à décrire K paramètres a_k. Pour cela, une étape importante est de pouvoir spécifier la précision avec laquelle il est nécessaire de coder chacun des paramètres a_k. Il faut se rappeler que les K paramètres a_k doivent être estimés à partir des N mesures (x_n, y_n). Le nombre N de mesures conditionne en particulier la précision avec laquelle les valeurs des a_k peuvent être connues (estimées), et donc la précision nécessaire au codage des a_k.

Pour l'estimation d'un paramètre a_k à partir de N mesures indépendantes, une limite fondamentale de la précision peut s'exprimer comme $1/\sqrt{NJ_1}$. Autrement dit, un nombre N de mesures ne permettent pas d'estimer la valeur du paramètre à mieux que $1/\sqrt{NJ_1}$ près. Ceci découle d'un résultat fondamental de la théorie statistique de l'estimation. La quantité $1/\sqrt{NJ_1}$ intervient dans une borne inférieure (la borne de Cramér-Rao [5, 2]), en dessous de laquelle l'erreur moyenne d'estimation ne peut pas descendre. La quantité J_1 représente ce qu'il est convenu d'appeler l'information de Fisher [5, 2] contenue dans une mesure au sujet du paramètre a_k.

La quantité NJ_1 représente l'information de Fisher contenue dans N mesures indépendantes au sujet du paramètre a_k. Plus l'information de Fisher est grande, plus les mesures permettent d'estimer précisemment le paramètre. En général, l'information de Fisher dépend du niveau du bruit ξ (elle décroît quand ξ augmente). Pour l'application du principe de longueur de description minimale, il n'est toutefois pas nécessaire de spécifier l'expression explicite de J_1, seule importe son augmentation en NJ_1 avec le nombre N de mesures.

Le codage d'un paramètre a_k mesuré avec la précision $1/\sqrt{NJ_1}$ peut être vu comme la localisation du paramètre a_k sur une échelle linéaire dont le pas de quantification est $\Delta q_N = 1/\sqrt{NJ_1}$. Si l'on note D_0 l'étendue du domaine où peut a priori se trouver a_k, avec un pas de quantification Δq_N on compte donc $D_0/\Delta q_N$ "cases" où peut se trouver a_k après N mesures. En particulier, avec une seule mesure ($N = 1$), on compte un nombre de cases de $D_0/\Delta q_1 = D_0\sqrt{J_1}$ où peut se trouver a_k. Avec N mesures, on compte $D_0/\Delta q_N = D_0\sqrt{NJ_1}$ cases où peut se trouver a_k. Il faut dans ce cas une longueur de description de $\log(D_0/\Delta q_N)$ pour coder ces $D_0/\Delta q_N$ cases possibles. S'il s'agit d'un code binaire par exemple, il faudra $\log_2(D_0/\Delta q_N)$ bits pour coder $D_0/\Delta q_N$ cases. La longueur de description nécessaire au codage du paramètre a_k peut donc s'évaluer comme $\log(D_0/\Delta q_N) = \log(D_0\sqrt{NJ_1})$ soit encore $\log(D_0\sqrt{J_1}) + \log(\sqrt{N})$.

Pour continuer, il est d'usage de procéder à une approximation [8]. Dans le présent contexte de l'estimation de paramètres inconnus à partir de N mesures bruitées, il est possible de négliger le terme $\log(D_0\sqrt{J_1})$ devant le terme $\log(\sqrt{N})$. Ceci se justifie car N est en général "grand". En même temps, le terme $D_0\sqrt{J_1}$ peut être vu comme proche de 1 car une unique mesure bruitée dans ce contexte ne permet pas en général de réduire beaucoup l'incertitude a priori sur le paramètre ; pour cela il faudra un nombre de mesures N grand.

Ainsi donc, lors de l'application du principe de longueur de description minimale, on compte comme $\log(\sqrt{N})$ la longueur de description nécessaire au codage d'un paramètre a_k. Pour coder K paramètres, il faudra K fois cette longueur, c'est-à-dire $K\log(\sqrt{N}) = L(\boldsymbol{a})$.

La longueur totale $L(\mathcal{M})$ pour décrire l'ensemble $\mathcal{M}$ des mesures peut donc s'exprimer comme

$$L(\mathcal{M}) = L(\mathcal{M}|\boldsymbol{a}) + L(\boldsymbol{a}) \,, \tag{11}$$

où $L(\mathcal{M}|\boldsymbol{a}) = -\log p(\mathcal{M}|\boldsymbol{a})$ est la longueur de description des mesures[1] lorsque les K paramètres $\boldsymbol{a}$ du modèle sont connus, et $L(\boldsymbol{a}) = K\log(\sqrt{N})$ est la longueur de description des K paramètres $\boldsymbol{a}$. Il vient donc

$$L(\mathcal{M}) = -\log p(\mathcal{M}|\boldsymbol{a}) + \frac{K}{2}\log(N) \,. \tag{12}$$

Le principe de longueur de description minimale prescrit alors de choisir $(\boldsymbol{a}, K)$, c'est-à-dire les valeurs des paramètres a_k ainsi que leur nombre K, de façon à minimiser $L(\mathcal{M})$ de l'Éq. (12). Les valeurs optimales $(\widehat{\boldsymbol{a}}_{\mathrm{LDM}}, \widehat{K}_{\mathrm{LDM}})$ qui résultent sont donc définies par

$$(\widehat{\boldsymbol{a}}_{\mathrm{LDM}}, \widehat{K}_{\mathrm{LDM}}) = \arg \min_{(\boldsymbol{a},K)} \; L(\mathcal{M}) \,, \tag{13}$$

soit

$$(\widehat{\boldsymbol{a}}_{\mathrm{LDM}}, \widehat{K}_{\mathrm{LDM}}) = \arg \min_{(\boldsymbol{a},K)} \left[-\log p(\mathcal{M}|\boldsymbol{a}) + \frac{K}{2}\log(N) \right] \,. \tag{14}$$

[1]Dans cet article, $p(\mathcal{M}|\boldsymbol{a})$ représente plus proprement une *densité* de probabilité ; la *probabilité* associée $P(\mathcal{M}|\boldsymbol{a})$ peut s'exprimer comme $P(\mathcal{M}|\boldsymbol{a}) = p(\mathcal{M}|\boldsymbol{a})\Delta V$ où ΔV est un volume élémentaire de l'espace des mesures. Toutefois, comme ΔV est choisi indépendamment des paramètres $\boldsymbol{a}$ et K du modèle à estimer, ΔV ne joue aucun rôle dans les critères à optimiser, ce qui permet de confondre la probabilité $P(\mathcal{M}|\boldsymbol{a})$ avec la densité $p(\mathcal{M}|\boldsymbol{a})$.

4 Exemples

Pour illustrer le principe de longueur de description minimale pour l'estimation de modèles paramétriques à partir de mesures bruitées, nous revenons à l'exemple de la régression polynomiale de l'Éq. (3). Dans le cas du bruit gaussien, sous les mêmes hypothèses qui ont conduit à l'Éq. (9), le principe de l'Éq. (13) conduit alors à

$$(\widehat{\boldsymbol{a}}_{\mathrm{LDM}}, \widehat{K}_{\mathrm{LDM}}) = \arg \min_{(\boldsymbol{a},K)} \left[\frac{1}{2\sigma^2} \sum_{n=1}^{N} \Big(y_n - Q_K(x_n) \Big)^2 + \frac{K}{2} \log(N) \right] . \qquad (15)$$

La Figure 1 fournit un programme informatique qui réalise la minimisation de l'Éq. (15). Ce programme est écrit pour le logiciel libre de calcul scientifique Scilab [9, 10]. Le programme de la Fig. 1, en testant différentes valeurs successives de l'ordre K, détermine les K paramètres a_k minimisant la fonction $\sum_{n=1}^{N}[y_n - Q_K(x_n)]^2/(2\sigma^2) + K\log(N)/2$ du membre de droite de l'Éq. (15). Cette fonction à minimiser est calculée sous le nom de `cout()` par le sous-programme de la Fig. 2 ; la minimisation elle-même est effectuée dans le programme principal de la Fig. 1 par l'instruction Scilab `optim()` qui appelle la fonction `cout()`. La valeur de K correspondant au plus petit minimum renvoyé définit la solution recherchée $(\widehat{\boldsymbol{a}}_{\mathrm{LDM}}, \widehat{K}_{\mathrm{LDM}})$.

```
a_vrai=[4, 5, 3];                    //vrais coefficients du modèle polynômial
sigma=10;                            //écart-type du bruit gaussien
Kmax=6;                              //ordre maximal qui sera testé pour le modèle
n=[1:1:100]; NN=length(n);           //n : les numéros des NN mesures
x=n/20;                              //x(n) : les abscisses des NN mesures
y_vrai=a_vrai(1)+a_vrai(2)*x+a_vrai(3)*x.^2;  //y_vrai : ordonnées non bruitées
y=y_vrai+sigma*rand(y_vrai, 'normal');        //y : ordonnées bruitées mesurées
plot2d(x, y);                        //graphe des NN couples (x_n, y_n) mesurés

getf('cout.sci');                    //chargement de la fonction 'cout' à minimiser

tab_a_opt=zeros(Kmax, Kmax);
for K=1:1:Kmax,
  a_init=ones([1:K]);                          //initialisation pour la minimisation
  [Lmin, a_opt]=optim(cout, a_init);           //minimisation de la fonction 'cout'
  tab_Lmin(K)=Lmin;
  tab_a_opt(K, 1:K)=a_opt;
end
tab_Lmin                             //affiche tableau des longueurs de description minimales

[Lmin, Kopt]=min(tab_Lmin);                    //Kopt : l'ordre optimal du modèle
a_opt=tab_a_opt(Kopt, 1:Kopt)                  //a_opt : les paramètres optimaux
poly_opt=poly(a_opt, 'x', 'coeff');            //création du polynôme
for i_n=1:1:NN,
  y_lis(i_n)=horner(poly_opt, x(i_n));         //évaluation du polynôme
end
plot2d(x, y_lis, style=[color('red')]);        //graphe du modèle optimal estimé
```

FIG. 1 – Programme principal en langage Scilab réalisant la minimisation de l'Éq. (15) déterminant $(\widehat{\boldsymbol{a}}_{\mathrm{LDM}}, \widehat{K}_{\mathrm{LDM}})$.

```
function [L, gradL, ind]=cout(a, ind)

  //Calcul de la longueur de description selon l'Éq. (15) :
  L=0;
  for n=1:1:NN,
    poly_degK=poly(a, 'x', 'coeff');
    Q_n=horner(poly_degK, x(n));
    e(n)=y(n)-Q_n;
    L=L+e(n)^2;
  end
  L=L/(2*sigma^2)+length(a)/2*log(NN);

  //Calcul du gradient de la longueur de description :
  for i_a=1:1:length(a),
    dcout_dak=0;
    for n=1:1:NN,
      dcout_dak=dcout_dak-2*e(n)*x(n)^(i_a-1);
    end
    gradL(i_a)=dcout_dak/(2*sigma^2);
  end
endfunction
```

FIG. 2 – Sous-programme en langage Scilab calculant la fonction `cout()` à minimiser $\sum_{n=1}^{N}[y_n - Q_K(x_n)]^2/(2\sigma^2) + K\log(N)/2$ du membre de droite de l'Éq. (15).

La Figure 3 présente les résultats de l'application du principe de longueur de description minimale, pour l'estimation d'un modèle d'ordre 3 selon la loi parabolique $Q_3(x) = a_1 + a_2 x + a_3 x^2$, à partir de $N = 100$ mesures bruitées. Le Tableau 1 donne les différentes valeurs de la longueur de description $\sum_{n=1}^{N}[y_n - Q_K(x_n)]^2/(2\sigma^2) + K\log(N)/2$, calculées par le programme de la Fig. 1 pour différentes valeurs de l'ordre K testées pour le modèle. Sur le Tableau 1, le minimum de la longueur de description est atteint en $K = 3$ qui estime correctement, de façon automatique, l'ordre du modèle.

ordre K du modèle	1	2	3	4	5	6
longueur de description	485.613	64.952	51.875	53.734	55.946	58.204

TAB. 1 – Longueur de description $\sum_{n=1}^{N}[y_n - Q_K(x_n)]^2/(2\sigma^2) + K\log(N)/2$ de l'Éq. (15), calculée par le programme de la Fig. 1 pour différentes valeurs de l'ordre K testées pour le modèle de la Fig. 3.

Les 3 paramètres (a_1, a_2, a_3) associés à l'ordre optimal $K = 3$ du modèle, sont aussi automatiquement estimés, avec une bonne précision compte-tenu du fort niveau de bruit, comme le montre la Fig. 3. Au sujet de cette précision de l'estimation, notons que, pour toute valeur de l'ordre K, les paramètres sont estimés au sens du maximum de vraisemblance. Par conséquent, la précision de l'estimation peut s'analyser statistiquement comme celle d'un estimateur du maximum de vraisemblance, profitant ainsi de ses propriétés théoriques générales et spécialement intéressantes [1, 2, 3]. En particulier, à N grand, l'estimateur est asymptotiquement sans biais et d'erreur quadratique moyenne d'estimation la plus faible parmi tous les estimateurs concevables (elle atteint la borne de Cramér-Rao déduite de l'information de

Fisher). Concrètement donc, au sens de ces critères statistiques, la méthode de la longueur de description minimale réalise la meilleure estimation des paramètres, et ce sur un modèle dont l'ordre est optimal.

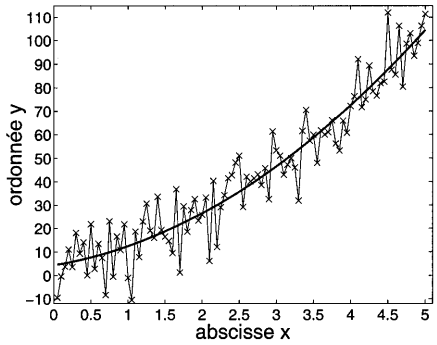

FIG. 3 – Les données synthétiques sont fabriquées à partir de la loi parabolique $Q_3(x) = a_1 + a_2x + a_3x^2$ avec $(a_1, a_2, a_3) = (4, 5, 3)$. Les croix ($\times$) sont les $N = 100$ points de mesure (x_n, y_n) avec les $N = 100$ abscisses x_n uniformément réparties sur l'intervalle $[0, 5]$ selon $x_n = [1 : 1 : 100]/20$ en présence de bruit ξ gaussien de moyenne nulle et d'écart-type $\sigma = 10$ selon l'Éq. (2). La courbe en trait épais est le modèle polynomial d'ordre optimal $\widehat{K}_{\text{LDM}} = 3$ et de coefficients optimaux $\widehat{\boldsymbol{a}}_{\text{LDM}} = (4.42, 5.12, 2.98)$ estimé à partir des N mesures selon le principe de longueur de description minimale.

Il importe de noter ici que le principe de longueur de description minimale n'est pas un procédé miraculeux ayant le pouvoir de découvrir les lois physiques cachées dans des mesures bruitées, et ce quelles que soient les conditions de mesure. Ce principe est plutôt un principe de modélisation de nature informationnelle, qui, à l'intérieur d'une classe prédéfinie, sélectionne le modèle le plus simple selon un critère informationnel, au vu d'un ensemble de mesures données, en tenant compte en particulier du niveau de bruit, du nombre de mesures, c'est-à-dire des conditions dans lesquelles les mesures ont été effectuées. À ce sujet, il est intéressant de considérer les deux exemples supplémentaires des Figs. 4–5. Il s'agit à chaque fois de l'estimation du même modèle en loi cubique $Q_4(x) = a_1 + a_2x + a_3x^2 + a_4x^3$, à partir de $N = 100$ mesures bruitées avec le même niveau de bruit.

Sur la Fig. 4, les $N = 100$ mesures sont prélevées dans une région où la loi cubique est plutôt linéaire, à l'échelle des fluctuations de bruit. Dans ce cas, le principe de longueur de description minimale sélectionne un modèle linéaire (voir Tableau 2), pour représenter économiquement ces mesures. Ce jeu de mesures, compte-tenu en particulier du niveau de bruit présent, ne contient pas assez d'information pour établir l'existence de la loi cubique sous-jacente.

ordre K du modèle	1	2	3	4	5	6	7
longueur de description	171.832	57.308	59.610	61.867	63.063	65.640	68.038

TAB. 2 – Longueur de description $\sum_{n=1}^{N}[y_n - Q_K(x_n)]^2/(2\sigma^2) + K\log(N)/2$ de l'Éq. (15), calculée par le programme de la Fig. 1 pour différentes valeurs de l'ordre K testées pour le modèle de la Fig. 4.

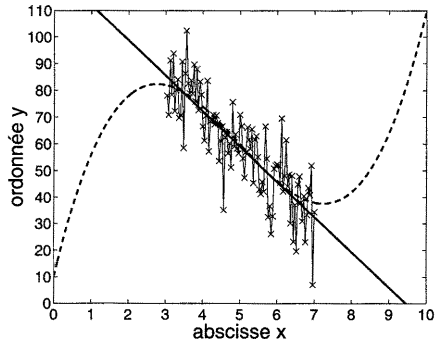

FIG. 4 – Les données synthétiques sont fabriquées à partir de la loi cubique $Q_4(x) = a_1 + a_2x + a_3x^2 + a_4x^3$ avec $(a_1, a_2, a_3, a_4) = (10, 60, -15, 1)$ représentée par la courbe en tirets. Les croix $(\times)$ sont les $N = 100$ points de mesure (x_n, y_n) avec les $N = 100$ abscisses x_n uniformément réparties sur l'intervalle $[3, 7]$ selon $x_n = 3 + 4[1 : 1 : 100]/100$ en présence de bruit ξ gaussien de moyenne nulle et d'écart-type $\sigma = 10$ selon l'Éq. (2). La droite en trait épais est le modèle polynomial d'ordre optimal $\widehat{K}_{\mathrm{LDM}} = 2$ et de coefficients optimaux $\widehat{\boldsymbol{a}}_{\mathrm{LDM}} = (125.20, -13.24)$ estimé à partir des N mesures selon le principe de longueur de description minimale.

Sur la Fig. 5, les $N = 100$ mesures sont prélevées dans une région mieux représentative de la loi cubique sous-jacente. Dans ce cas, le principe de longueur de description minimale sélectionne bien un modèle cubique (voir Tableau 3), pour représenter économiquement ces mesures, avec une bonne estimation des 4 paramètres de la loi cubique, compte-tenu du fort niveau de bruit, comme le montre la Fig. 5. À l'examen visuel, le jeu de mesures de la Fig. 5 peut apparaître essentiellement concave (∪), et un modèle parabolique aurait pu être suggéré pour lui. Le principe de longueur de description minimale indique néanmoins qu'un modèle cubique est mieux approprié. Surtout, ce principe dispense de ce genre de considérations visuelles subjectives, qu'il remplace par une approche quantitative objective et automatisable pour le choix du modèle.

ordre K du modèle	1	2	3	4	5	6	7
longueur de description	193.022	89.815	75.142	55.883	58.153	59.363	62.318

TAB. 3 – Longueur de description $\sum_{n=1}^{N}[y_n - Q_K(x_n)]^2/(2\sigma^2) + K\log(N)/2$ de l'Éq. (15), calculée par le programme de la Fig. 1 pour différentes valeurs de l'ordre K testées pour le modèle de la Fig. 5.

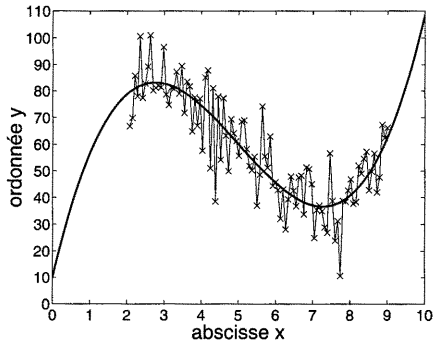

FIG. 5 – Les données synthétiques sont fabriquées comme sur la Fig. 4 à partir de la même loi cubique $Q_4(x) = a_1 + a_2x + a_3x^2 + a_4x^3$ avec $(a_1, a_2, a_3, a_4) = (10, 60, -15, 1)$. Les croix (×) sont les $N = 100$ points de mesure (x_n, y_n) avec les $N = 100$ abscisses x_n uniformément réparties sur l'intervalle $[2, 9]$ selon $x_n = 2 + 7[1 : 1 : 100]/100$ en présence de bruit ξ gaussien de moyenne nulle et d'écart-type $\sigma = 10$ selon l'Éq. (2). La courbe en trait épais est le modèle polynomial d'ordre optimal $\widehat{K}_{\mathrm{LDM}} = 4$ et de coefficients optimaux $\widehat{\boldsymbol{a}}_{\mathrm{LDM}} = (9.92, 60.87, -15.24, 1.01)$ estimé à partir des N mesures selon le principe de longueur de description minimale.

5 Conclusion

Nous avons exposé le principe de longueur de description minimale pour la modélisation des données à partir de mesures bruitées. Cette méthode, basée sur un critère informationnel,

permet, pour une classe quelconque de modèles spécifiée à nombre variable de paramètres, d'estimer à la fois le nombre optimal de paramètres ainsi que leurs valeurs optimales. D'un point de vue pratique, la méthode opère via la minimisation de l'Éq. (14). Comme on l'a dit, l'expression de l'Éq. (14) pour la longueur de description est valide, et conduit en général à de bons résultats pratiques, lorsque le nombre N de mesures est assez grand, ce qui est usuellemment la règle dans ce contexte d'estimation à partir de mesures bruitées. Cette méthode d'estimation de modèle qui s'appuie sur un critère quantitatif objectif, est complètement automatisable sous forme de programmation informatique, comme nous l'avons illustré à travers quelques exemples. En pratique, la formule de l'Éq. (14) et les programmes des Figs. 1–2 permettent d'aborder une très large variété de problèmes de modélisation de données.

Le principe de longueur de description minimale est par ailleurs applicable à d'autres situations d'inférence statistique, ou d'extraction optimale d'information dans un contexte statistique. Il peut revêtir des formes diverses, avec des expressions distinctes pour la longueur de description. Ce principe possède aussi des propriétés théoriques profondes, ainsi que des applications pratiques variées, les unes et les autres étant encore actuellement en développement [11]. L'ensemble permet de faire progresser un point de vue quantitatif formalisé au sujet de l'information associée aux mesures physiques.

Références

[1] J. L. Féménias, *Probabilités et Statistiques pour les Sciences Physiques.* Paris : Dunod, 2003.

[2] P. Réfrégier, *Théorie du Bruit et Applications en Physique.* Paris : Hermes, 2002.

[3] T. Alhalel, "Incertitudes et mesure des incertitudes," *Le BUP, Bulletin de l'Union des Professeurs de Physique et de Chimie*, n° 879 (2), déc. 2005, vol. 99, pp. 7–35.

[4] G. Battail, *Théorie de l'Information.* Paris : Masson, 1997.

[5] T. M. Cover, J. A. Thomas, *Elements of Information Theory.* New York : Wiley, 1991.

[6] J. C. Culioli, *Introduction à l'Optimisation.* Paris : Ellipses, 1994.

[7] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[8] J. Rissanen, *Stochastic Complexity in Statistical Inquiry.* Singapore : World Scientific, 1989.

[9] www.scilab.org

[10] J. P. Chancelier, F. Delebecque, C. Gomez, M. Goursat, R. Nikoukhah, S. Steer, *Introduction à Scilab.* Berlin : Springer, 2001.

[11] P. Grünwald, I. J. Myung, M. Pitt, *Advances in Minimum Description Length : Theory and Applications.* Cambridge (MA) : MIT Press, 2005.