Théorie de l'Information
Notes de Cours (part 5)
2006-2007

SIC-SICOM

Maria-João Rendas

19 novembre 2006

Table des matières

Chapitre 6

Identification de modèles : le principe de longueur de description minimale (MDL)

6.1 Introduction

Le problème d'identification de modèles peut être formulé comme celui de choisir une parmi plusieurs explications alternatives pour un ensemble de données $x^{(n)}$ (en nombre n limité, en toute situation pratique). Dans ce Chapitre, nous exposons une méthode pour l'identification de modèles connue par le nom de "*principe de description minimale*" ou encore par l'acronyme **MDL** (de l'anglais *Minimum Description Length*). Ses premières versions ont été proposées par N. Rissanen dans les années 1970 [2], et bâtissent sur les travaux d'autres chercheurssur la *théorie de la compléxité* (notamment le chercheur russe Kolmogorov [3]) qui datent des années 60. Cette méthodologie est basée sur deux idées fondamentales :

1. *Apprendre un modèle* pour le système qui a engendré les données observées consiste à détecter les *régularités* présentes dans l'ensemble observé. L'identification de modèles est une tâche d'*apprentissage*, où l'objectif est de capturer tant que possible les *régularités des données*, et, donc, être capable de trouver sa description la plus simple : séparer ce qui est *structure* de ce qui est *accidentel* (bruit).
2. Les *régularités* des données (l'ensemble de règles, ou propriétés, qui cet ensemble satisfait) peuvent être exploitées pour les *compréSSer*, c'est à dire, pour les décrire (exactement) avec un nombre de symboles minimal (inférieur à celui qui décrit les observations). Un modèle correspond donc à un *langage* pour décrire les données. Dans le cas de modèles décrits par des *familles de distributions de probabilité* (hypothèses composées dans la terminologie des tests de décision statistique), un bon modèle doit correspondre à un *codeur universel* pour les données dans la famille considérée.

Ces deux observations restent valables indépendament de considérations sur l'existence d'un *vrai modèle probabiliste* qui aurait engendré les données observées. L'approche du MDL se distingue d'une manière fondamentale, sur ce point particulier, des approches statistiques classiques où ce "vrai" modèle (probabiliste) est un ingrédient fondamental (comme l'approche Bayesienne, où la distribution *a priori* est censée traduire notre connaissance sur l'identité de ce *vrai* modèle). Les différentes distributions dans un même modèle correspondent à des différentes réalisations des mêmes régularités structurelles. Bien que n'excluant pas le type de modèles "signal + bruit", pour le MDL le bruit est le nombre de bits nécessaires pour identifier *la* séquence observée à l'aide d'un modèle qui contraint ses régularités. Un modèle avec un niveau de bruit élevé implique uniquement que les données ne sont pas compressibles avec le modèle considéré. Pour le MDL, l'inférence est basée *uniquement sur les données observées*.

Une procédure d'inférence statistique est *consistante* si elle identifie la *vraie* hypothèse avec probabilité 1 (asymptotiquement, pour un nombre de données n très grand). Bien que le MDL ne dépende pas, comme nous venons de le dire, de considérations sur l'existance de cette *vraie distribution*, il est important que, si jamais les données sont une réalisation d'une des distributions contenues dans les modèles considérés, cette distribution soit correctement identifiée. Comme nous le verrons, le *MDL est consistant*.

Comme nous le verrons avec plus de détail le long de ce Chapitre, la méthode MDL trouve des analogies et relations avec d'autres méthodes et problèmes d'analyse de données :

- **Rasoir d'Occam**. Le "rasoir" d'Occam (ou *principe de parsimonie*), énoncé par le philosophe médiéval William Occam (siècle XVI, Angleterre), sous la forme "pluritas non est ponenda sine necessitate," (une pluralité ne doit pas être invoquée sans besoin), conduit, dans le cas du problème d'identification de modèles, à choisir *le modèle le plus simple qui n'est pas contredit par les données*. Comme nous le verrons plus tard, le principe du MDL conduit lui aussi à la recherche d'un équilibre entre la *complexité* du modèle qui est ajusté aux données, et sa *capacité* pour les décrire exactement. En fait, quand deux modèles s'ajustent également bien aux données, MDL choisit le plus simple (dans le sens qu'il permet description plus courte), et dans ce sens, il réalise une sorte de "rasoir d'Occam". Nous remarquons cependant que cette préférence n'est pas un choix "phylosophique", qui impliquerait une quelconque supposition sur la simplicité de la nature, mais uniquement la traduction du fait qu'il n'y a pas de sens à identifier des modèles complexes avec un volume de données réduit. En fait, comme nous le verrons dans une section de ce Chapî tre, pour une même source le modèle sélectionné par MDL peut être de plus en plus complexe, à mesure que plus de données sont observées.
- Le problème de **sur-ajustement** (*overfitting*, en anglais). Ce problème est bien connu dans les problèmes d'identification de modèles paramétriques, où l'**ordre du modèle**, c'est à dire, le nombre de degrés de liberté du modèle qui est utilisé pour décrire les observations, est inconnu et doit être aussi déterminé à partir des données. Le MDL donne une réponse formellement justifiée (sur les proprié-

FIG. 6.1 – William Occam.

tés des codes universels) pour le choix de l'ordre du modèle. Il conduit ainsi à un équilibre entre les capacités de *description* des données et les propriétés de *généralisation* du modèle identifié.

– **Approche Bayesienne**. L'approche Bayesienne (du nom du révérend anglais Thomas Bayes, XVIII-ème siècle), fourni une méthode systématique pour *combiner (fusionner) des informations imprécises*. Dans le cadre du problème d'identification de modèles, l'approche Bayesienne conduit aussi à la recherche d'un équilibre entre la plausabilité pour que les données soient engendrées par le modèle, et sa probabilité *a priori*. Une des objections qui sont soulevées à l'approche Bayesienne est l'arbitrarité du choix de ces distributions *a priori*. Nous pouvons essayer de contre-carrer le problème de sur-ajustement en associant des probabilités *a priori* plus grandes aux modèles plus simples (ce que le "rasoir" d'Occam semble sugérer : "la nature préfère la simplicité"). Comme nous le verrons, le MDL possède des relations étroites avec l'approche Bayesienne, et peut être interprétré comme une méthodologie pour choisir la distribution *a priori*.
– **Codage prédictif**. Plusieurs méthodes d'identification ont été proposées pour lesquelles le critère pour la sélection du modèle est la capacité prédictive du modèle, c'est à dire, pour prédire des valeurs qui n'ont pas encore été observées. Comme nous le discuterons plus tard, le MDL possède aussi des relations avec

FIG. 6.2 – Thomas Bayes.

ce critère.

Ce Chapître est organisé de la façon suivante. Dans la Section 6.2 nous motivons le problème d'appretissage d'un modèle pour un ensemble d'observations comme une recherche du langage (code) qui les décrit de la manière la plus compacte. Dans la section 6.3 nous rappelons les relations étroites entre lois de probablité et codes optimaux qui sont à la base du MDL. La section 6.4 présente une version particulière du MDL (codage en deux parties) qui permet une interprétation intuitive de son fonctionnement. La section 6.5 présente la notion de codeur universel, et propose une formulation générale du MDL basée sur cette notion, qui nous conduira à la définition de *complexité stochastique*. La section 6.6 présente des expressions asymptotiques pour la complexité de modèles paramétriques, mettant en évidence la relation du MDL avec d'autres approches au problème d'apprentissage de modèles. Finalement, la section 6.7 discute des généralisations de la définition précédante, nottamment pour le cas de modèles pour lesquels la complexité paramétrique est infinie.

6.2 Apprentissage (de modèles) comme compression de données

Pour établir d'une manière intuitive la relation entre le problème d'identification d'un modèle pour un ensemble de données et le problème de compression de données, nous considérons les trois séquences binaires du tableau suivant

	séquence
x_1	001001001001001001001001001001001001001001001001001001
x_2	011101011001110000001111011101111000100110010111000011001
x_3	000000000100101000100100001000000110100010000000000000001

La séquence x_1 est la répétition du motif '001'. Elle a une *structure très simple* et peut être décrite d'une manière efficace par la règle (programme) suivante

Écrire 20 fois la séquence '001'. Stop.

Cette règle est elle même une suite de symboles (caractères) qui peuvent être codés avec des symboles binaires (par exemple, avec le code ASCII).
La séquence x_2 est une suite de 56 tirages indépendants d'une loi uniforme en $\{0, 1\}$. Elle ne possède donc pas de régularités. Comme nous avons vu dans les Chapîtres précédants, la séquence x_2 ne peut pas être comprimée car son entropie (en bits) est égale à sa longueur.
La séquence x_3 est "moins aléatoire" que la séquence x_2. Nous pouvons comencer par constater que le nombre de 0's dans cette séquence (45) est beaucoup plus grand que le nombre de 1's (11). Nous pouvons exploiter ce fait pour coder efficacement x_3. En fait, il n'existe qu'un nombre fini (et beaucoup plus petit que 2^{56} !) de séquence binaires avec 11 éléments égales à '1'. Il est possible d'ordonner (par exemple par ordre lexicographique) ces séquences. Un code plus efficace pour x_3 commencerait donc par coder le nombre k de symboles '1' dans x_3 ($k = 11$) et ensuite indiquer uniquement l'index $i_k(x_3)$ de la séquence x_3 dans l'ensemble de toutes les séquences de taille 56

avec k '1"s :

$$x_3 \rightarrow (k, i_k(x_3))$$

Un exemple d'une nature différente (et plus près de beaucoup de cas practiques de problèmes d'identification de modèle), et posé par le codage des valeurs du tableau suivant

x_i	y_i
-5	-286.4260
-4	-160.7825
-3	-85.7549
-2	-26.8838
-1	3.9952
0	16.7044
1	7.0395
2	7.0602
3	1.1894
4	-5.0388
5	8.2898

La figure 6.3 représente y_i en fonction de x_i. La forme de cette courbe sugère une dépendance polynomiale de y_i en x_i :

$$y_i = P_k(x_i) = a_k x_i^k + a_{k-1} x_i^{k-1} + \cdots + a_1 x_i + a_0$$

Nous pouvons alors coder efficacement les données (y_i, x_i) en envoyant d'abord l'odre k du polynome $P_k(\cdot)$ et ses coefficients $\{a_n\}_{n=0}^k$ (ou ses zéros) suivis de mots de code pour les valeurs de $\{x_i\}_{i=1}^n$ et pour les résidus $\{\epsilon_i\}_{i=1}^n$ du modèle polynomial (différence entre la valeur du poynome $P(x_i)$ et les valeurs y_i) :

$$\{(y_i, x_i)\}_{i=1}^n \rightarrow \left(k, \{a_n\}_{n=0}^k, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n\right)$$

Si, $n \gg 1$ et $k \ll n$ cette description des données doit être plus compacte que la description directce des données : nous avons exploité la relation (polynomiale) sous-jacente à tous les pairs pour les décrire plus efficacement

Code de préfixe pour les entiers

Dans l'exemple précédant nous avons eut besoin (pour coder les séquences x_1 et x_2, de coder des entiers. Nous allons maintenant introduire un **code (de préfixe) pour les entiers** qui sera utilisé par la suite.

Pour toute séquence binaire $x \in \{0,1\}^\star$, (de taille $\ell(x) \geq 1$), nous pouvons construire un code de préfixe pour x, en le faisant précéder le code binaire de x par un mot $C_{sd}(m)$ qui indique la taille m de la séquence binaire ultérieurement utilisée pour coder la valeur de x :

$$x \rightarrow C(x) = C_{sd}(m) C_m(x)$$

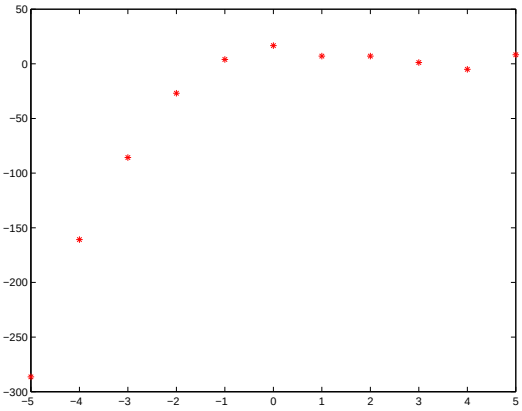

FIG. 6.3 – William Occam.

où $C_{sd}(m)$ est un code *auto-délimité* (*self-delimited*) pour l'entier m, et $C_m(x)$ est un code de longueur constante égale à m.
Le préfixe $C_{sd}(m)$ peut être, par exemple, une séquence de $m = \ell(x)$ zéros suivie d'un 1 :

$$x = 0100110; \qquad \ell(x) = 7 \qquad \Rightarrow C(x) = 000000010100110,$$

où $m = 7$, $C_{sd}(7) = 00000001$ et $C_7(x) = 0100110$.
Le décodage de $C(x)$ est simple : nous identifions la fin du préfixe par l'occurrence du symbole '1'. Nous pouvons alors déterminer la taille m de la séquence qui code x, et donc, l'extraire du message.
Considérons maintenant le codage d'un entier $k \in \{1, 2, 3, \ldots\}$. Comme $k \in \{1, 2^{\lceil \log k \rceil}\}$, nous pouvons coder k avec un code où le préfixe $C_{sd}(m)$ est une séquence de $m = \lceil \log k \rceil$ zéros suivie d'un 1.
Le nombre total de bits de ce code de préfixe pour les entiers est égal à $2\lceil \log k \rceil + 1$.
Par exemple, pour $k = 21$, le code sera

$$00000\ 1\ 10101$$

Le récepteur comence par conter le nombre de '0's : 5, il sait donc que les 5 bits après le '1' contiennent le code pour k.

6.3 Codes et lois de probabilité

Le lien entre le choix d'hypothèses/modèles et la compréssion des données n'a pas encore été établi. De l'étude du problème de codage source que nous avons menée dans les Chapitres précédants, nous savons que

1. à un code (de préfixe, complet) C sur un alphabet $\mathcal{X}$ nous pouvons associer une loi de probabilité p_C sur $\mathcal{X}$:

$$C \rightarrow p_C(x) = 2^{-\ell(c(x))}, \forall x \in \mathcal{X}.$$

où $\ell(c(x))$ est la longueur du mot de code que C associe à $x \in \mathcal{X}$. L'inégalité de Kraft nous affirme que p_C est en général (pour des codes qui ne sont pas complets) une distribution non-normalisée (une *semi-mesure*), avec une masse totale inférieure ou égale à 1. Si le code est *complet*, alors, p_C est une loi de probabilité. Nous savons encore, du théorème de Shannon du codage sans pertes, que le code C est optimal (il conduit à une longueur moyenne de code minimale, égale à l'entropie de la source $H(X)$) :

$$L_C(X) = \mathrm{E}\left[\ell(c(x))\right] = H(X).$$

2. à une loi de probabilité $p(x)$,$x \in \mathcal{X}$, nous pouvons associer un code optimal C_p (de longueur moyenne minimale), avec des longueurs de mots de code

$$p(x) \rightarrow \ell(C_p(x)) = \lceil -\log p(x) \rceil.$$

Ce code (de Shannon-Fano) a une redondance inférieure à 1 bit. En fait, de la non-négativité de l'entropie relative, il découle que

$$\mathrm{E}_p\left[-\log Q(x)\right] \geq \mathrm{E}_p\left[-\log p(x)\right].$$

(*inégalité fondamentale de l'information*.)

Nous introduisons maintenant la terminologie précise qui sera utilisée dans tout ce Chapître.

Définition 1 *Modèle probabiliste*
Um modèle probabiliste $\mathcal{H}$ est un *ensemble* de sources décrites par lois de probabilité. Les modèles *paramétriques*, où les éléments de l'ensemble $\mathcal{H}$ sont indéxés par un vecteur (de dimension finie) de paramètres θ appartenant à un espace de paramètres Θ est particulièrement intéressant :

$$\mathcal{H} = \{p(X|\theta) \ : \ \theta \in \Theta\}, \qquad \Theta \subset \Re^k, k \geq 1.$$

△

Exemple 1 *Modèle Gaussien*
Pour des observations $x^{(n)} \in \Re^n$, le modèle Gaussien

$$\mathcal{G} = \left\{p(x^{(n)}|\theta) \ : \ \theta = (\mu, \Sigma)\right\}, \qquad \mu \in \Re^n, \qquad \Sigma = \Sigma^T \geq 0 \in \Re^{n \times n},$$

paramétré par la moyenne μ et la matrice de covariance Σ, où

$$p(x^{(n)}|\theta) = (2\pi \det(\Sigma))^{-1/2} \exp\{-\frac{1}{2}(x^{(n)} - \mu)^T \Sigma^{-1}(x^{(n)} - \mu)\},$$

est un modèle paramétrique, paramétré par $n + n(n+1)/2$ paramètres (où nous avons considéré les conditions de symmétrie de la matrice de covariance. △

Définition 2 *Estimateur du Maximum de Vraisemblance*
Pour un modèle probabiliste $\mathcal{H}$ et des données $x^{(n)} \in \mathcal{X}^{(n)}$, l'estimateur du Maximum de Vraisemblance est l'élément du modèle qui maximise la probabilité des données :

$$\hat{p}_{MV}(x^{(n)}) = \arg\max_{p \in \mathcal{H}} p(x^{(n)}).$$

Pour les modèles paramétriques, ce modèle est obtenue en prennant l'élément de $\mathcal{H}$ qui est indéxé par l'estimateur du Maximum de Vraisemblance des [aramètres θ :

$$\hat{p}_{MV}(x^{(n)}) = p(\cdot|\hat{\theta}_{MV}(x^{(n)})), \qquad \hat{\theta}_{MV}(x^{(n)}) = \arg\max_{\theta \in \Theta} p(x^{(n)}|\theta).$$

△

Définition 3 *Estimateur consistant*
Nous disons qu'un estimateur est consistant si, pour des observations

$$x^{(n)} \sim p_n \rightarrow_{n\to\infty} p^\star, \qquad x^{(n)} \in \mathcal{X}^n$$

nous avons

$$\lim_{n\to\infty} \hat{p}_{MV}(x^{(n)}) = p^\star,$$

où la convergence est en probabilité. Pour les modèles paramétriques, où $p^\star = p(\cdot|\theta^\star)$,

$$\lim_{n\to\infty} \hat{\theta}_{MV}(x^{(n)}) = \theta^\star.$$

△

Nottez que la consistance est une propriété *asymptotique*, elle concerne le comportement des estimateurs pour un grand nombre n de données : avec probabilité 1, l'estimateur tend vers la vraie mesure qui a engendré les données.

Exemple 2 *Modèle de Markov*
Nous avons défini dans un Chapitre précédant les chaînes de Markov d'ordre 1 (invariantes dans le temps) avec état $x_n \in \mathcal{X}$, où $|\mathcal{X}| = m$. Ce modèle est paramétré par la distribution initiale $p(x_0)$ ($m-1$ paramètres) et par la matrice de transition $P_{ij} = p(x_{n+1} = i|x_n = j)$ ($m(m-1)$ paramètres). La dimension de ce modèle est donc $(m+1) \times (m-1)$.
Nus introduisons maintenant les *modèles de Markov d'ordre k*, où la valeur à l'instant n dépend des k valeurs précédantes :

$$p(X_n = i|x_1^{n-1}) = p\left(X_n = i|x_{n-k}^{n-1}\right), \qquad n > k.$$

Dans ce cas, le modèle est indéxé par sa distribution initiale $p(x_1^k)$ avec $k \times (m-1)$ paramètres, et par une matrice de transition de dimension $(m^k - 1) \times (m^k - 1)$.
Le modèle de Markov *binaire* ($m = 2$) d'ordre k avec $\mathcal{X} = \{0, 1\}$, demande donc un

total de $k + (2^k - 1)^2$ paramètres.
Il peut être vérifié facilement que pour une chaîne de Markov d'ordre $k = 1$

$$\left[\hat{P}_{MV}\right]_{ij}(x^{(n)}) = \frac{n_{ij}}{n}, \tag{6.1}$$

où n_{ij} est le nombre de fois que le symbole i est précédé par le symbole j dans la séquence $x^{(n)}$. △

Définition 4 *Modèle de Bernoulli*
Le modèle de Bernoulli (valeurs binaires statistiquement indépendantes et identiquement distribuées), que nous notterons $\mathcal{B}$, peut être obtenu comme un modèle de Markov binaire d'ordre $k = 0$, avec une matrice de transition

$$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}.$$

Il est paramétré par un seul paramètre $\theta \in [0, 1]$, qui est la probabilité d'occurrence d'une des valeurs de $\mathcal{X}$:

$$\mathcal{B} = \left\{ p(x^{(n)}|\theta) \ : \ \theta \in [0,1] \right\}, \qquad p(x^{(n)}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^{n_1}(1-\theta)^{n-n_1},$$

où n_1 est le nombre de '1's dans la séquence $x^{(n)}$. Il est aisé de vérifier que pour le modèle de Bernoulli

$$\hat{\theta}_{MV}(x^{(n)}) = \frac{n_1}{n}.$$

△

6.4 MDL (codage en deux parties)

6.4.1 Complexité de Kolmogorov MDL "Idéal"

Les exemples que nous avons présenté dans l'Introduction construisent des codes efficaces pour les messages considérés, mais ces codes sont trouvés, à chaque cas, par des méthodes *ad hoc*. Pour que l'approche MDL soit bien définie, il faut spécifier les types de codage qui sont des candidats pour coder les données.
Dans les années 60, **Kolmogorov** [3] a proposé la définition d'une mesure de compléxité *intrinsèque* d'un objet X comme la *taille du plus petit programme qui serait capable de produire* X (une séquence, une image, une forme,...) *et s'arrêter*. Il a appelé cette mesure de complexité (que l'on peut démontrer qui est indépendante du langage de programation choisi) *complexité algorithmique*. Dans le cadre de cette théorie (développée parallèlement par Solomonoff [4] et Chaitin [5]), les données sont décrites (codées) par un programme (qui doit s'exécuter sur une machine de Turing). La notion de complexité de Kolmogorov a depuis été reprise par plusieurs chercheurs, et est

d'une grande puissance formelle. Cependant, il est possible de démontrer qu'il n'existe pas un programme qui puisse la déterminer : la complexité de Kolmogorov n'est pas *calculable*, ce qui ne permet pas la considérer comme la base d'une méthode (pratique) d'identification.
Néanmoins, certains auteurs [6] ont proposé une version "idéale" du principe MDL, en considérant le choix du modèle dans lequel la compléxité de Kolmogorov serait la plus petite.

Nous allons nous intéresser à des versions "pratiques" du principe du MDL, et donc à des méthodes de description qui soient *calculables*. Le prix à payer sera que nous devrons abandonner les langages de programation génériques, pour passer à considérer la complexité d'un message par rapport à un ensemble de langages (une "encyclopédie"), et la complexité du message ne sera plus une propriété intrinsèque du message. Certaines séquences régulières ne seront pas comprimées.

Nous énonçons maintenant le principe du MDL d'une manière informelle, et présenterons une approche (codage en deux parties) qui, sous certaines condtions, en est une implémentation.

6.4.2 Codage en deux parties

Définition 5 *Principe du MDL (définition informelle)*
Soit $\mathcal{M}$ un ensemble d'hypothèses (de modèles probabilistes $\mathcal{H}$), et $x^{(n)} \in \mathcal{X}^n$ des données. Le principe MDL nous dit de choisir le modèle $\mathcal{H}^\star \in \mathcal{M}$ qui *comprime le plus les données.*

$$\mathcal{H}^\star = \arg \min_{\mathcal{H} \in \mathcal{M}} \min_{H \in \mathcal{H}} L_H(x^{(n)}),$$

où $L_H(x)$ est la longueur du mot d'un *code optimal pour les données x correspondant au modèle H.* △

Les deux faits que nous avons établi dans la Section précédante justifient l'association de l'ensemble d'hypothèses probabilistes $\mathcal{H}$ dans la définition 5 avec un ensemble de lois de probabilité (et donc de codes). La définition suivante précise une interprétation possible de la notion de "codage optimal avec un modèle" utilisée dans la définition informelle 5.

Définition 6 *MDL (codage en deux parties)*
Soient $\mathcal{H}^{(i)}$, $i = 1, 2, \ldots$, un ensemble de modèles, et $x^{(n)}$ des données. Alors le *meilleur modèle pour les données*, $\mathcal{H}^\star$ est le plus petit modèle qui contient

$$H^\star = \arg \min_{H \in \cup_{i=1,2,\ldots} \mathcal{H}^{(i)}} L(H) + L(x^{(n)}|H), \tag{6.2}$$

où $L(H)$ est la taille nécéssaire pour décrire l'hypothèse, et $L(x^{(n)}|H)$ le nombre de bits nécessaire pour coder les données davec l'hypothèse (code) H. △

Le premier terme dans l'équation (6.2) est la longueur de description de l'hypothèse H. Le deuxième terme est la longueur de description des données $x^{(n)}$ à l'aide de l'hypothèse H, et est une mesure de l'*ajuste du modèle aux données*. Remarquons que cette version en deux parties du MDL rend explicite l'idée de chercher un équilibre entre la complexité du modèle utilisé (modèles plus complexes demandront un $L(H)$ plus grand), et l'ajuste des modèles aux données $x^{(n)}$ (plus les données seront plausibles (probables) dans le modèle, plus petite sera la longueur de son code).

La définition précédante laisse sans réponse deux questions :

1. Comment détermine-t-on $L(H)$?
2. Comment calculer $L(x^{(n)}|H)$?

La réponse à la dernière question, le *choix de* $L(x^{(n)}|H)$, est relativement simple : si nous admettons qu'à chaque hypothèse H correspond une loi de probabilité $p(x^{(n)}|H)$ (car $\mathcal{H}$ est un modèle probabiliste), alors nous devons prendre comme longueur de description des données la taille du code optimal :

$$L(x^{(n)}|H) = -\log p(x^{(n)}|H).$$

La longueur des mots de ce code est égale à moins la vraisemblance des données – ce qui correspond bien à une mesure d'ajustement. Nous savons que, si les données suivent la distribution $p(x^{(n)}|H^\star)$, alors pour ce choix la taille $L(x^{(n)}|H)$ sera minimisée pour la bonne hypothèse $H = H^\star$. Si (comme nous allons le voir) $L(H)$ est indépendante de n (ou si $L)H)$ croit sub-linéairement en n), alors, pour des grandes valeurs de n la somme dans l'équation (6.2) sera dominée par $L(x^{(n)}|H)$. La propriété d'équi-répartition asymptotique nous garanti alors que si les données sont une réalisation d'une des hypothèses contenues dans un des modèles, alors, la méthode MDL identifiera ce modèle avec probabilité 1 (elle est donc consistante).

La réponse à la première question, *le choix de* $L(H)$, est moins simple : si nous choisissons arbitrairement la façon dont les hypothèses sont codées, nous pouvons, avec des schémas différents, avoir des codes qui atribuent des tailles très différentes aux hypothèses, et nous n'aurions donc pas une façon objective de choisir le modèle. Comme nous le verrons, ce problème a conduit à la proposition de différentes versions alternatives du principe du MDL.

Souvent, l'ensemble de modèles a une structure *emboitée*

$$\mathcal{M} = \cup_{k>0}\mathcal{H}^{(k)},$$

où la dimension de $\mathcal{H}^{(k)}$ croit avec k, et

$$\mathcal{H}^{(k)} \subset \mathcal{H}^{(k+1)}, k = 1, 2, \ldots$$

C'est le cas, par example, quand chaque $\mathcal{H}^{(k)}$ est le modèle d'observations bruitées de l'ensemble de polynomes d'ordre k ou pour l'ensemble modèles de Markov d'ordre k. Dans ces cas, pour chaque valeur de k nous pouvons utiliser un code fixe, en deux parties, pour coder les différentes hypothèses contenues dans $\mathcal{H}^{(k)}$, comme c'est illustré par l'exemple suivant.

Exemple 3 *Un test du MDL pour les modèles de Markov*
Nous pouvons décrire un modèle de Markov (binaire) d'ordre k en indiquant en premier lieu la valeur de k, suivie de la valeur du paramètre $\theta \in [0,1]^{2^k}$:

$$H \in \mathcal{H}^{(k)} \rightarrow (k, \theta).$$

L'entier k peut être codé avec le code présenté page 100, qui demande $2\lceil \log k \rceil + 1$ bits :

$$L(k) = 2\lceil \log k \rceil + 1.$$

Pour coder le paramètre θ, nous devons indiquer les valeurs des 2^k probabilités p_{1i}, pour que le symbole 1 soit précédé par la i-ème séquence s_i de k symboles binaires, $s_i \in \{0,1\}^k$, (cela détermine la probabilité pour que le symbole '0' soit précédé par la même séquence : $p_{0i} = 1 - p_{1i}$).
Nous remarquons maintenant que nous pouvons restreindre l'ensemble $\mathcal{H}^{(k)}$ de tous les modèles de Markov d'ordre k ùn sous-ensemble discret fini $\mathcal{H}^{(k)'}$. En fait, uniquement les modèles qui peuvent être identifiés à partir d'une séquence de taille n sont pertinents, ce qui implique que les probabilités p_{1i} soient de la forme de l'équation (6.1), page 103, où, nécessairement

$$n_{ij} \in \{0, \ldots, n\}.$$

Construire un code qui associerait des mots de code à des hypothèses qui ne correspondent pas à des estimés possibles aurait conduit à une valeur supérieure pour $L(H)$ sans que les valeurs correspondantes de $L(x^{(n)}|H)$ puissent être plus petites.

Il n'y a donc que $n+1$ valeurs possibles pour chaque probabilité p_{1i} (fréquence n_{1j}), qui peuvent être codés avec un nombre de bits non supérieur à $\lceil \log(n+1) \rceil \simeq \log(n+1)$. Comme le paramètre θ du modèle de Markov est défini par 2^k de ces valeurs, pour identifier un modèle particulier nous avons besoin d'un nombre de bits total égal à

$$L(H) = L(k) + L(\theta) = 2\log k + 1 + 2^k \log(n+1).$$

Le critère pour choisir le modèle de Markov (l'ordre et les paramètres) est donc la solution de

$$\min_k \min_\theta - \log p(x^{(n)}|\theta) + 2\log k + 1 + 2^k \log(k+1),$$

ou encore, en faisant appel à la définition d'estimateur de Maximum de Vraisemblance

$$\min_k - \log p(x^{(n)}|\hat{\theta}_{MV}) + 2\log k + 1 + 2^k \log(k+1).$$

$\triangle$

6.5 Codeurs universels et MDL (codage en une partie)

La Section précédante a illustré l'application du principe du MDL avec une approche de codage en deux parties : codage de l'hyopthèse (avec $L(H)$ bits) suivie du codage des données dans l'hypothèse (avec $-\log \hat{p}_{\mathcal{H}}(x^{(n)}$ bits).
Cependant, le choix du codage effectué est heuristique, et n'identifie pas d'une façon formelle le code qui doit être associée à un modèle probabiliste $\mathcal{H}$. Nous allons introduire la notion *codeur universel dans une classe de modèles*, qui permet de dépasser ce problème.

Définition 7 *Codeur universel (dans une classe de modèles)*
Soit $\mathcal{H}$ un modèle probabiliste (et donc un ensemble de longeurs de codes optimaux) pour des données $x^{(n)}$. Un code $\overline{L}$ est universel pour $\mathcal{H}$ s'il est capable (assymptotiquement) de coder la séquence $x^{(n)}$ avec un nombre de bits égal au code optimal pour $x^{(n)}$ dans le modèle $\mathcal{H}$:

$$\forall x^{(n)} \qquad \lim_{n\to\infty} \frac{1}{n}\overline{L}(x^{(n)}) = \lim_{n\to\infty} \min_{H\in\mathcal{H}} \frac{1}{n} L(x^{(n)}|H). \tag{6.3}$$

Nottons que si la séquence $x^{(n)}$ est une réalisation d'une des hypothèses dans $\mathcal{H}$, alors ce limite est égal au taux d'entropie $\overline{H}(X)$. △

Nous remarquons que cette définition de code universel correspond à une notion de code universel *ponctuel*, c'est à dire, la propriété (6.3)est vérifiée pour toutes les séquences $x^{(n)}$. Dans le Chapître précédant nous avons présenté, dans le contexte de létude de l'algorithme de Lempel-Ziv, une notion de code universel qui est basée sur la longueur *moyenne* des mots de code.

Les deux exemples suivants montrent que *les codes en deux parties sont des codes universels*.

Exemple 4 *Nombre fini d'hypothèses*
Soit $x^{(n)}$ une séquence binaire, et considérons le modèle suivant :

$$\mathcal{H} = \{\mathcal{B}(\theta_i), i = 1, \dots 9\}, \qquad \theta_i = \frac{i}{10},$$

de façon que

$$-\log p(x^{(n)}|\mathcal{B}(\theta_i)) = -n_1 \log(i/10) - (n - n_1) \log\left(\frac{10-i}{10}\right).$$

Nous pouvons construire un code universel (en deux parties) pour ce modèle de la façon suivante. Nous codons d'abord $i^\star$,

$$i^\star = \arg \max_{i=1,\dots,9} p(x^{(n)}|\mathcal{B}(\theta_i))$$

en utilisant un code uniforme en $\{1,\ldots,9\}$, avec donc $\lceil \log 9 \rceil$ bits. Le message $x^{(n)}$ est ensuite codé avec le code $\mathcal{B}_{\theta_{i^\star}}$, ce qui demande

$$L_{i^\star} = -\log p\left(x^{(n)}|\mathcal{B}(\theta_{i^\star})\right) = L_{\mathcal{B}(\theta_{i^\star})}(x^{(n)})$$

bits. Ce code a une taille totale

$$\overline{L}_{\mathcal{H}}(x^{(n)}) = \lceil \log 9 \rceil - \log p\left(x^{(n)}|\mathcal{B}(\theta_{i^\star})\right) = \lceil \log 9 \rceil + \min_{H\in\mathcal{H}} L_H(x^{(n)}).$$

Nous pouvons facilement constater que

$$\lim_{n\to\infty} \frac{1}{n}\overline{L}(x^{(n)}) = \lim_{n\to\infty} \frac{1}{n} \min_{H\in\mathcal{H}} L_H(x^{(n)}),$$

qui est la limite atteinte par le meilleur code pour $x^{(n)}$ dans le modèle $\mathcal{H}$, et que le code présenté est donc universel pour le modèle considéré. △

Dans cet exemple nous avons utilisé un *code uniforme* pour coder les éléments $H \in \mathcal{H}$ (l'index $i^\star$ de l'exemple précédant), ce qui correspond à admettre une distribution uniforme pour toutes les hypothèses H du modèle $\mathcal{H}$. Autres choix seraient possibles. Cependant, ce choix minimise la redondance du code dans le *pire cas* (le cas où la vraie hypothèse reçoit le mot de code le plus long pour la distribution admise).

Exemple 5 *Modèle avec un ensemble dénombrable d'hypothèses*
Considérons maintenant le cas d'un modèle avec un nombre infini d'hypothèses :

$$\mathcal{H} = \{H_1, H_2, \ldots\}.$$

Un codage en deux parties pour ce modèle peut être obtenu, d'une façon analogue à l'exemple précédant, en considérant le codage de l'index $i^\star$ avec le code pour les entiers que nous avons introduit page 100, qui demande $2\lceil \log k \rceil + 1$ bits pour coder l'entier k. Nous devons donc maintenant choisir

$$i^\star = \arg \inf_{i=1,2,\ldots} \left\{2\lceil \log i \rceil + 1 + L_{H_i}(x^{(n)})\right\},$$

et donc

$$\overline{L}(x^{(n)}) = \inf_{i=1,2,\ldots} \left\{2\lceil \log i \rceil + 1 + L_{H_i}(x^{(n)})\right\}.$$

Contrairement au cas précédant, nous ne pouvons plus borner par une constante la redondance de ce code par rapport au code qui atteint la longueur de code minimale dans le modèle $\mathcal{H}$. Par contre,

$$\lim_{n\to\infty} \frac{1}{n}\overline{L}(x^{(n)}) = \lim_{n\to\infty} \inf_{i=1,2,\ldots} \left\{\frac{2\log i + 1}{n} - \frac{1}{n}\log p(x^{(n)}|H_i)\right\},$$

et, si les observations $x^{(n)}$ sont une réalisation d'une des lois de probabilité $p(\cdot|H_{i_0})$[1] dans $\mathcal{H}$,

$$\lim_{n\to\infty} \frac{1}{n}\overline{L}(x^{(n)}) = \overline{H}(p_{i^0})$$

[1] et si le modèle correspondant est ergodique.

où $\overline{H}(p_{i^0})$ est le taux d'entropie associé à la loi de probabilité correspondante à la vraie distribution des données. Pour cet exemple, le taux avec lequel le nombre de bits par symbole approche le taux d'entropie de la source n'est plus constant comme pour l'exemple précédant. $\triangle$

L'association entre codes et lois de probabilité nous permet d'associer aux codes utilisés pour décrire le modèle (avec des longueurs de code $L(H)$) des lois de probabilité définies dans l'ensemble des hypothèses contenues dans $\mathcal{H}$

$$p(H) = 2^{-L(H)}, \qquad H \in \mathcal{H},$$

établissant ainsi un pont entre le codage MDL en deux parties et l'approche Bayesienne. L'exemple suivant illustre cette relation, montrant que *les marginales de Bayes sont aussi des modèles (codes) universels*.

Exemple 6 *Modèle Bayesien universel*
Soit $\mathcal{H}$ un ensemble fini ou dénombrable d'hypothèses (lois de probabilté), paramétrées par un vecteur de paramètres $\theta \in \Theta$:

$$\mathcal{H} = \left\{ p(x^{(n)}|\theta) \ : \ \theta \in \Theta \right\}.$$

Soit W une distribution de probabilité dans Θ. À chaque distribution W nous pouvons associer un *modèle de mélange* pour les observations :

$$p(x^{(n)}|\mathcal{H}, W) = \sum_{\theta \in \Theta} W(\theta) p(x^{(n)}|\theta). \tag{6.4}$$

Il est immédiat que $p(x^{(n)}|\mathcal{H}, W)$ est un modèle universel pour $\mathcal{H}$. La longueur du code correspondant est :

$$\overline{L}_{\mathcal{H},W}(x^{(n)}) = -\log p(x^{(n)}|\mathcal{H}, W).$$

Et donc

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \overline{L}_{\mathcal{H},W}(x^{(n)}) &\overset{(a)}{\leq} -\lim_{n\to\infty} \frac{1}{n} \log W(\theta) p(x^{(n)}|\theta) \\ &\leq -\lim_{n\to\infty} \frac{1}{n} \max_{\theta} \left(\log W(\theta) + \log p(x^{(n)}|\theta) \right) \end{aligned} \tag{6.5}$$

où (a) est vraie pour tout choix de θ dans le membre droit. Ceci découle du fait que tous les termes de la somme sont non-négatifs, et de la monotonicitté du logarithme. Si nous nottons θ_0 la valeur de θ correspondante à ce maximum,

$$\lim_{n\to\infty} \frac{1}{n} \overline{L}_{\mathcal{H},W}(x^{(n)}) \leq -\lim_{n\to\infty} \frac{1}{n} \log p(x^{(n)}|\theta_0) = \overline{H}(p(\cdot|\theta_0)) \ ,$$

car, parce que $W(\theta_0)$ est une constante indépendante de n,

$$\lim_{n\to\infty} \frac{1}{n} \log W(\theta_0) = 0.$$

Ceci démontre que la distribution de mélange Bayesienne de l'équation (6.4) conduit à un code universel pour les données. Elle est donc un exemple de *modèle universel* (pour le modèle $\mathcal{H}$). Une analyse de l'équation (6.5) nous montre que le code universel de Bayes est supérieur au codage en deux parties avec un code $L(H)$ dérivé de la distribution *a priori* W. Dans ce dernier cas, nous serions conduits à une longueur de code

$$-\min_{\theta} \log p(x^{(n)}|\theta) + \log W(\theta),$$

qui, sauf pour le cas où $p(x^{(n)}|\theta) = 0, \forall \theta \neq \theta_0$, est strictement supérieur à la valeur dans le membre gauche de (6.5), correspondante au modèle de mélange. △

6.5.1 Maximum de Vraisemblance Normalisé comme Modèle Universel Optimal

Définition 8 *Pénalité* $\mathcal{P}_{\overline{P},\mathcal{H},x^{(n)}}$

Soit $\mathcal{H}$ un modèle probabiliste, et soit $\overline{P}$ une loi de probabilité définie en $\mathcal{X}^n$. Soient $x^{(n)}$ les observations, $x^{(n)} \in \mathcal{X}^n$.

La *pénalité* de la loi $\overline{P}$ par rapport au modèle $\mathcal{H}$ pour les observations $x^{(n)}$ est, par définition

$$\mathcal{P}_{\overline{P},\mathcal{H},x^{(n)}} = -\log \overline{P}(x^{(n)}) - \min_{p \in \mathcal{H}} \left(-\log p(x^{(n)}) \right).$$

Cette pénalité est la différence entre *(i)* la taille du code associé à $\overline{P}$, et le *(ii)* nombre de bits nécessaire pour coder les observations *avec le meilleur code dans le modèle* $\mathcal{H}$. Pour le cas de modèles paramétriques, $\mathcal{H} = \{p(\cdot|\theta) \ : \ \theta \in \Theta\}$,[2]

$$\mathcal{P}_{\overline{P},\mathcal{H},x^{(n)}} = -\log \overline{P}(x^{(n))}) + \log p\left(x^{(n)} | \hat{\theta}_{MV}(x^{(n)}) \right). \tag{6.6}$$

Avec cette définition, nous pouvons déjà constater qu'une loi $\overline{P}$ sera un modèle (code) universel pour le modèle $\mathcal{H}$ si pour toute séquence $x^{(n)}$ la pénalité croit plus lentement que n

$$\forall x^{(n)} \qquad \lim_{n \to \infty} \frac{1}{n} \mathcal{P}_{\overline{P},\mathcal{H},x^{(n)}} = 0.$$

La notion de pénalité, qui dépend de la séquence observée $x^{(n)}$, ne doit pas être confondue avec celle de *redondance*, introduite dans les Chapitres précédants, et qui concerne le comportement *en moyenne* de la longueur des mots du code.

Il est intéressant d'étudier la pénalité pour un modèle (code) universel pour $\mathcal{H}$. Cependant, la valeur de la pénalité, telle qu'elle est proposée dans la définition précédante, dépend de la séquence observée $x^{(n)}$. Pour certaines séquences elle peut même être négative, et en même temps qu'elle prend des valeurs grandes pour d'autres séquences. La définition suivante considère la plus grande valeur (sur toutes les observations $x^{(n)}$)

[2] Nous admettons ici que l'estimée du Maximum de Vraisemblance est bien définie.

de la pénalité pour un code $\overline{P}$ et un modèle $\mathcal{H}$.

Définition 9 *Pénalité au pire cas* $\mathcal{R}_{\overline{P},\mathcal{H}}$
Soit $\mathcal{H}$ un modèle probabiliste, et $\overline{P}$ une loi de probabilité en $\mathcal{X}^n$. La *pénalité au pire cas* de $\overline{P}$ par rapport à $\mathcal{H}$ est, par définition

$$\mathcal{R}_{\overline{P},\mathcal{H}} = \max_{x^{(n)} \in \mathcal{X}^n} \mathcal{P}_{\overline{P},\mathcal{H},x^{(n)}}. \tag{6.7}$$

Nous sommes en ce moment en position de définir un **modèle universel optimal** $\overline{P}^\star$ comme celui qui *minimize la pénalité au pire cas* :

$$\begin{aligned} \overline{P}^\star &= \arg\min_{\overline{P}} \mathcal{R}_{\overline{P},\mathcal{H}} \\ &= \arg\min_{\overline{P}} \max_{x^{(n)} \in \mathcal{X}^n} \left\{ -\log \overline{P}(x^{(n)}) + \log p\left(x^{(n)} | \hat{\theta}_{MV}(x^{(n)})\right) \right\}, \end{aligned}$$

où nous avons considéré le cas d'un modèle probabiliste paramétrique. Ce problème d'optimisation a une solution qui est égale à

$$\overline{P}_{mvn}(x^{(n)}) = \frac{p\left(x^{(n)} | \hat{\theta}_{MV}(x^{(n)})\right)}{\sum_{y^{(n)} \in \mathcal{X}^n} p\left(y^{(n)} | \hat{\theta}_{MV}(y^{(n)})\right)}, \tag{6.8}$$

quand le dénominateur de cette expression est fini.

La distribution (6.8) est connue par le nom de *distribution de Shtarkov*, du nom du chercheur qui a établi son optimalité. Elle associe à chaque possible séquence observée $x^{(n)}$ une probabilité qui est proportionnelle à celle qui lui est assignée par le modèle correspondant à l'estimateur du Maximum de Vraisemblance pour $x^{(n)}$ (la distribution qui lui attribue la probabilité la plus grande). Ceci explique la notation "mvn" dans l'expression précédante, indiquant "*Maximum de Vraisemblance Normalisé*".

La distribution définie en (6.8) est toujours bien définie quand $\mathcal{H}$ est un ensemble *fini*. Autrement, la somme dans le dénominateur peut être infinie, et donc $\overline{P}_{mvn}$ n'est pas défini.

Définition 10 *Complexité paramétrique* $\mathcal{C}(\mathcal{H})$
Soit $\mathcal{H}$ un modèle probabiliste paramétrique. Sa *complexité* est, par définition

$$\mathcal{C}(\mathcal{H}) = \log \sum_{x^{(n)} \in \mathcal{X}^n} p\left(x^{(n)} | \hat{\theta}_{MV}(x^{(n)})\right). \tag{6.9}$$

Avec cette définition, la distribution de Shtarkov s'écrit

$$\log \overline{P}_{mvn}(x^{(n)}) = \log p\left(x^{(n)} | \hat{\theta}_{MV}(x^{(n)})\right) - \mathcal{C}(\mathcal{H}).$$

De par sa définition, nous pouvons constater que la compléxité paramétrique d'un modèle, $\mathcal{C}(\mathcal{H})$ est d'*autant plus grande que le modèle $\mathcal{H}$ peut décrire* (avec probabilité élevée) *un grand nombre de séquences* $x^{(n)}$. Comme nous le verrons, la complexité stochastique $\mathcal{C}(\mathcal{M})$ est liée au nombre de degrés de liberté du modèle $\mathcal{M}$.

La démonstration de l'optimalité de $\overline{P}_{mvn}$ est immédiate. Si nous utilisons (6.8) dans la définition (6.6), nous obtenons

$$\mathcal{R}_{\overline{P}_{mvn},\mathcal{H}} = \log \sum_{x^{(n)} \in \mathcal{X}^n} p\left(x^{(n)} | \hat{\theta}_{MV}(x^{(n)})\right) = \mathcal{C}(\mathcal{H}),$$

indépendamment de $x^{(n)}$. La pénalité devient donc indépendante de la séquence $x^{(n)}$. Comme $\forall p \neq \overline{P}_{mvn}$ il doit exister au moins un $z^{(n)} \in \mathcal{X}^n$ tel que $p(z^{(n)}) < \overline{P}_{mvn}(z^{(n)})$,

$$\begin{aligned}
\mathcal{R}_{p,\mathcal{H}} &= \max_{x^{(n)} \in \mathcal{X}^n} \mathcal{P}_{p,\mathcal{H},x^{(n)}} \\
&\geq \mathcal{P}_{p,\mathcal{H},z^{(n)}} = -\log p(z^{(n)})) + \log p\left(z^{(n)} | \hat{\theta}_{MV}(z^{(n)})\right) \\
&> -\log \overline{P}_{mvn}(z^{(n)})) + \log p\left(z^{(n)} | \hat{\theta}_{MV}(z^{(n)}\right) \\
&= \mathcal{R}_{\overline{P}_{mvn},\mathcal{H},z^{(n)}} = \mathcal{R}_{\overline{P}_{mvn},\mathcal{H}}.
\end{aligned}$$

Ceci montre que la pénalité au pire cas de toutes les autres distributions p dans $\mathcal{H}$ doit être supérieure à celle de $\overline{P}_{mvn}$.

6.5.2 MDL (codage en un partie) et complexité stochastique

Nous présentons maintenant une formulaiton du principe de la longueur de description minimale, qui fait appel à la notion de modèle universel optimal présentée dans la section précédante, et qui permet de résoudre quelques problèmes associés aux codes en deux parties étudiés dans la Section 6.4.

Nous conduisons cette présentation dans le cadre simple de choix entre deux modèles $\mathcal{H}_1$ et $\mathcal{H}_2$, et en admettant que la *complexité de ces deux modèles est finie*, de façon que les *distributions de Shtarkov correspondantes*, que nous nottons $\overline{P}_{mvn}(\cdot|\mathcal{H}_i)$, $i = 1, 2$, *sont bien définies*. Nous reviendrons dans une section postérieure sur le problème de choisir entre un nombre infini de modèles.

Définition 11 *Principe MDL (codage en une partie)*
Soient $\mathcal{H}_1, \mathcal{H}_2$ des modèles alternatifs pour les observations $x^{(n)}$, et $\overline{P}_{mvn}(\cdot|\mathcal{H}_i), i = 1, 2$, les distributions qui minimizent la pénalité au pire cas par rapport aux modèles $\mathcal{H}_i$, $i = 1, 2$, respectivement. Alors, le principe du MDL nous dit de choisir le modèle $\mathcal{H}_{j^\star}$ pour lequel la longueur du mot de code qui est associé aux observations par le modèle universel optimal est minimale :

$$j^\star = \arg\min_{j=1,2} -\log \overline{P}_{mvn}\left(x^{(n)} | \mathcal{H}_j\right) \tag{6.10}$$

$$= \arg\min_{j=1,2} \left\{ -\log p\left(x^{(n)} | \hat{\theta}_j(x^{(n)})\right) + \mathcal{C}\left(\mathcal{H}_j\right) \right\}. \tag{6.11}$$

△

Cette dernière expression nous donne une interprétation en deux parties de cette version reformulée du MDL : le modèle optimal $\mathcal{H}_{j^\star}$ réalise un équilibre entre la capacité de décrire exactement les données (le terme $p\left(x^{(n)}|\hat{\theta}_j(x^{(n)})\right)$) et sa complexité paramétrique (mesurée par le terme $\mathcal{C}\left(\mathcal{H}_j\right)$).

La définition précédante du principe du MDL implique donc que le modèle choisit pour les données est celui pur lequel le code universel a le meilleur comportement (pénalité minimale) dans *le pire cas*.

Définition 12 *Complexité stochastique des données* $\overline{L}(x^{(n)}, \mathcal{H})$
Soit $\mathcal{H}$ un modèle probabiliste (paramétrique). La complexité des observations $x^{(n)}$ pa rapport au modèle $\mathcal{H}$ est, par définition, la probabilité qui lui est attribuée par la distribution de Shtarkov associée à $\mathcal{H}$:

$$\overline{L}(x^{(n)}, \mathcal{H}) = -\log \overline{P}_{mvn}\left(x^{(n)}|\mathcal{H}\right).$$

△

Le principe du MDL que nous venons de formuler, nous indique donc de prendre *le modèle pour lequel la complexité stochastique des observations est minimale*.

En général, la complexité paramétrique d'un modèle (et donc la complexité stochastique des observations dans ce modèle) ne peut pas être calculée, ni même numériquement, la seule exception connue étant le cas Gaussien. Cependant, des approximations de la complexité stochastique des données $\overline{L}(x^{(n)}, \mathcal{H})$, valables pour n grand, peuvent être déterminées, comme c'est présenté dans la section suivante.

6.6 Approximations de la complexité stochastique

6.6.1 Maximum de Vraisemblance Généralisé

Pour les problèmes de décision entre plusieurs hypothèses composées $\mathcal{H}_i$, (où chaque hypothèse est associée à une *famille* de lois de probabilité), le principe du Maximum de Vraisemblance Généralisé nous dit de choisir l'hypothèse $\mathcal{H}_{i^\star}$ qui maximize

$$p_{\hat{p}_i}(x^{(n)}) = \max_{p \in \mathcal{H}_i} p(x^{(n)}) + c,$$

où c est une constante qui détermine la performance du test (la probabilité des différents types d'erreur). Souvent, une version simplifiée de ces tests est appliquée, qui consiste à ignorer la constante c et à choisir simplement l'hypothèse qui maximise la probabilité des données. Ceci correspond, dans la perspective du MDL que nous venons d'énoncer dans la section précédante, à négliger la complexité paramétrique de chaque modèle, $\mathcal{C}(\mathcal{H}_i)$, et pour des modèles avec un nombre de degrés de liberté disfférents, en général

conduit à un choix systématique du modèle le *plus complexe* (celui qui a le plus grand nombre de degrés de liberté, et qui peut donc décrire plus précisément les observations $x^{(n)}$).
Le MDL essaie de contrarier cette tendance pour choisir un modèle de complexité élevée en considérant la distribution de Shtarkov, qui utlilise, comme nous avons vu, une version *normalisée* du Maximum de Vraisemblance. Le plus grand sera le nombre d'observations qui peuvent être bien décrites par les distributions contenues dans un modèle, plus grande sera sa complexité paramétrique $\mathcal{C}(\mathcal{H})$, et donc plus le modèle doit finement décrire les observations $x^{(n)}$ pour qu'il puisse être choisi.

6.6.2 MDL et Compression

L'idée originale de Rissanen qui justifie le principe du MDL est de considérer le problème d'identification de modèles comme celui d'apprendre le modèle qui le mieux exprime les régularités présentes dans les données. Nous venons de voir que cela correspond à choisir le modèle pour lequel la distribution de Shtarkov associée conduit à un code de longueur minimale. L'utilisation de $\overline{P}_{mvn}$ semble justifiée par les deux remarques suivantes :

1. le mieux la meilleure distribution dans $\mathcal{H}_j$ décrira les données, le plus petite sera la longueur de code.
2. aucune distribution dans chaque $\mathcal{H}_j$ n'est donnée une préférence : en effet, la pénalité (qui a le sens d'une redondance *pour chaque séquence*) est la même pour toutes les séquences $x^{(n)}$, indépendemment de la loi qui a générée la séquence observée. Ce code est le seul code de préfixe avec cette propriété : $\overline{P}_{mvn}$ traite toutes les distributions de la même manière.

La complexité stochastique des données $x^{(n)}$ dans un modèle $\mathcal{H}$ peut être interprétée comme la *quantité d'information sur le modèle contenue dans les données*. Elle est la somme de deux termes. La complexité paramétrique du modèle, qui mesure la quantité de *structure* dans les données (dans le contexte du modèle $\mathcal{H}$), et le terme $-\log p(x^{(n)}|\hat{\theta}(x^{(n)})$, qui mesure la quantité de *bruit* dans les données.

6.6.3 Interprétation géométrique

La complexité paramétrique d'un modèle peut être interprétée comme le nombre de distributions *discernibles* dans le modèle. Intuitivement, comme nous l'avons déjà remarqué, plus de distributions différentes un modèle contient plus grand est le nombre de messages qu'il peut décrire, et donc plus grand est le risque de over-fitting. Cependant, si ces distributions sont "similaires", dans le sens qu'elles décrivent essentiellement le même ensemble d'observations, elles ne doivent pas contribuer séparament à la complexité stochastique du modèle. Dans cette perspective, la complexité devrait être une mesure du nombre de distributions "différentes" qu'un modèle donné contient. L'analyse suivante montre dans un cas simple que $\mathcal{C}(\mathcal{H})$ mesure exactement ceci.

Soit $\mathcal{H} = \{p(\cdot|\theta_i), i \in \{1, \ldots, M\}\}$ un modèle avec un nombre fini de distribu-

tions. Alors

$$\begin{aligned}
\mathcal{C}(\mathcal{H}) &= \log \sum_{x^{(n)} \in \mathcal{X}^n} p(x^{(n)}|\hat{\theta}_{MV}(x^{(n)}) \\
&= \log \sum_{j=1,\ldots,M} \sum_{x^{(n)} \,:\, \hat{\theta}_{MV}(x^{(n)}=\theta_j} p(x^{(n)}|\theta_j) \\
&= \log \sum_{j=1,\ldots,M} \left(1 - \sum_{x^{(n)} \,:\, \hat{\theta}_{MV}(x^{(n)} \neq \theta_j} p(x^{(n)}|\theta_j)\right) \\
&= \log \left(M - \sum_{j=1,\ldots,M} \Pr\left\{\hat{\theta}_{MV}(x^{(n)} \neq \theta_j | \theta_j\right\}\right) \\
&\leq \log M \; .
\end{aligned}$$

Nous voyons donc que la complexité paramétrique est dans ce cas simple égale à la différence entre le nombre de modèles dans $\mathcal{H}$ (M) et la probabilté pour que les modèles soient confondus. L'existence de distributions qui peuvent être confondues avec une probabilité élevée conduit à une diminution de la complexité paramétrique (ces distributions ne sont pas comptabilisées d'une façon indépendante). Pour des grandes valeurs de n, cette probabilté d'erreur tend vers zéro (sauf dans des cas pathologiques), et la complexité paramétrique tend vers $\log M$, le (logarithme du) nombre de modèles différentes dans $\mathcal{H}$.

Ce fait a été formellement établi en [1], à l'aide d'arguments sur la géométrie de la variété Riemanienne associée à des modèles probabilites paramétriques $\{p(x^{(n)}|\theta) \; : \theta \in \Theta\}$ par la définition d'une métrique qui est donnée par la matrice de Fisher $I(\theta)$. (On admet que Θ est compact.) Nous présentons ici très sommairement l'argument utilisé. Considérons les ellipsoides centrés sur une *grille discrète de points* $\theta_i \in \Theta$:

$$(\theta - \theta_i)^T I(\theta_i)(\theta - \theta_i) \leq d(n)$$

où la grille $\{\theta_i\}$ est telle que les rectangles maximales contenus dans ces ellipsoides déterminent un partitionnent non-uniforme de l'espace de paramètres Θ. Considérons maintenant que $d(n)$ dépend du nombre de données n de telle façon que le *volume de ces ellipsoides* peut être approximé par

$$V_i(n) = |\det(I(\theta_i))|^{-1/2} \left(\frac{2\pi}{n}\right)^{k/2},$$

de façon que si θ_i et θ_j sont des éléments *voisins* dans cette grille, alors, la probablité d'erreur

$$\Pr\{\hat{\theta}(x^{(n)}) = \theta_i | \theta_j\} \rightarrow_{n\to\infty} 0, \qquad i \neq j.$$

Le taux de diminution du volume $V_i(n)$ des ellipsoides (et donc d'augmentation du nombre de points dans la grille discrète θ_i) est critique, dans le sens que pour une variation de $d(n)$ qui induirait une diminution plus rapide de $V_i(n)$, la probablité de confondre deux points voisins de la grille ne convergerait pas vers zéro. L'auteur montre

que le *nombre de points dans la grille* ainsi construite est égal à la complexité paramétrique $\mathcal{C}(\mathcal{H})$, qu'il désigne par *complexité géométrique*. Nous pouvons donc comprendre le terme correspondant à la complexité paramétrique dans le code universel optimal de Shtarkov (maximum de vraisemblance normalisé) comme la longueur des mots de code pour une distribution uniforme $p(\theta_i) = 1/\mathcal{C}(\mathcal{H})$ dans cet ensemble de *modèles discernibles* en $\mathcal{H}$ pour les observations $x^{(n)}$. Le terme $-\log p(x^{(n)}|\hat{\theta}(x^{(n)})$, code les données comme une des séquences qui correspondent à l'élément de la partition associée au modèle choisit, c'est à dire, les "détails" des données.

Remarque 1
L'analyse que nous venons de présenter montre que pour des modèles finis les codes en deux parties avec une distribution a priori *uniforme sont asymptotiquement optimales. Comme nous l'avons vu (exemple 4) la pénalité de ces codes est* $\log M$*, et nous venons de voir que pour le code universel optimal la pénalité prend asymptiquement la valeur de* $\log M$*. Cependant, pour des valeurs de* n *petits, et pour certaines séquences* $x^{(n)}$*, la probabilité d'erreur est non-nulle, et donc la complexité paramétrique est inférieure à cette valeur asymptotique. Le code en deux parties considère donc une pénalité qui sera pour certaines séquences, supérieure à celle du code optimal en une seule partie.* △

Pour le cas plus intéresssant de modèles qui contiennent un nombre infini de distributions, la complexité paramétrique a une interprétation comme un *quotient* de deux volumes. Cette interpétation est basée dans l'expansion asymptotique de la distribution $\overline{P}_{mvn}$ présentée dans le lemme suivant.

Lemme 1
Soit $\mathcal{H}$ un modèle paramétrique de dimension m. Sous certaines conditions de régularité

$$\mathcal{C}(\mathcal{H}) = \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\theta\in\Theta}\sqrt{\det(I(\theta))}\,d\theta + o(1), \tag{6.12}$$

où n est la longueur de la séquence observée, $I(\theta)$ est la matrice de Fisher pour le vecteur de paramètres θ, d'élément générique

$$[I(\theta)]_{ij} = \lim_{n\to\infty}\frac{1}{n}\mathrm{E}_\theta\left\{-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x^{(n)}|\theta)\right\},$$

et $o(1) \to 0$ quand $n \to \infty$. △

Le dernier terme en (6.12) ne dépend pas de n et donc sa contribution devient négligeable pour des valeurs de n grandes, impliquant que pour n très grand,

$$\mathcal{C}(\mathcal{H}) \simeq \frac{k}{2}\log\frac{n}{2\pi}, \qquad n >> 1. \tag{6.13}$$

Le membre droit de cette expression coincide (à part un facteur $k/2\log 2\pi$) avec le terme de pénalité utilisé par la méthode BIC (Bayesian Information Criterion) pour le

probléme d'identification de modèles. Le fait que MDL et BIC, pour des valeurs suffisament grandes de n, conduisent à la même pénalité a conduit à l'affirmation (erronée) que MDL et BIC sont équivalents.

En particulier, le Lemme 1 est vrai si $\mathcal{H}$ est une famille exponentielle :

$$p(x^{(n)}|\theta) = \exp(\beta^T t(x^{(n)}))f(x^{(n)})g(\beta).$$

Cette famille contient un grand nombre de densités usuelles, comme les Bernoulli et multinomiale Gaussienne, Gamma, etc.

Le premier terme de (6.12) mesure directement le nombre de degrés de liberté du modèle, k. Le deuxième terme est une correction qui dépend de la forme fonctionnelle du modèle. C'est une correction qui ne dépend pas de n, et qui donc peut être négligée pour des valeurs de n très grandes.

Exemple 7 *Complexité du modèle de Bernoulli*
Pour le modèle de Bernoulli introduit dans la page 103 la dimension est $k = 1$. La matrice de Fisher pour ce modèle est

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

L'utilisation de ce résultat dans l'expansion asymptotique (6.12) conduit à

$$\mathcal{C}(\mathcal{B}) = \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} - 3 + o(1).$$

△

6.6.4 Interprétation Bayesienne

Les tests de décision Bayesiens ont un lien fort avec la méthode MDL. Considérons le test entre deux hypothèses composées suivant

$$\begin{aligned} H_1 \quad &: \quad x^{(n)} \sim p_1(x) \in \mathcal{H}_\infty = \left\{p(x^{(n)}|\theta^1), \theta^1 \in \Theta^1\right\}, \theta^1 \sim w^1(\theta^1) \\ H_2 \quad &: \quad x^{(n)} \sim p_2(x) \in \mathcal{H}_\in = \left\{p(x^{(n)}|\theta^2), \theta^2 \in \Theta^2\right\}, \theta^2 \sim w^2(\theta^2) \end{aligned}$$

où les hypothèses sont équiprobables : $\Pr\{\mathcal{H}_1\} = \Pr\{\mathcal{H}_2\}$. Le test optimal au sens de Bayes (pour le critère de probabilité d'erreur minimale) choisit l'hypothèse $i^\star$ qui a la *plus grande probabilité a posteriori* :

$$i^\star = \arg\max_{i\in\{1,2\}} \overline{p}(x^{(n)}|\mathcal{H}_i), \qquad \overline{p}(x^{(n)}|\mathcal{H}_i) = \int_{\Theta^i} p(x^{(n)}|\theta^i)w^i(\theta^i)\,d\theta^i. \quad (6.14)$$

Quand les $\mathcal{H}_i$ sont des **familles exponentielles**, et sous conditions de régularité, une expansion de Laplace (dévelopement en série de la fonction intégrée dans l'expression

précédante) conduit à l'approximation suivante

$$-\log \overline{p}\left(x^{(n)}|\mathcal{H}_i\right) = -\log p\left(x^{(n)}|\hat{\theta}^i(x^{(n)})\right) + \frac{k}{2}\log\frac{n}{2\pi} - \log w\left(\hat{\theta}^i(x^{(n)})\right) + \log\sqrt{\det(I(\theta^i(x^{(n)}))} + o(1).$$

Si nous comparons cette expression à l'équation (6.12), nous pouvons constater que la longueur de code atteinte par ce code "de Bayes" *diffère par une constante* de la longueur optimale $-\log \overline{P}_{mvn}$. Pour des grandes valeurs de n, les deux approches conduiront donc au choix du même modèle.

Si nous considérons le cas particulier où les distributions *a priori* sont les distributions de Jeffrey ("least informative prior", introduite par Jeffrey en 1946) :

$$w(\theta) = \frac{\sqrt{det(I(\theta))}}{\int_{\theta' \in \Theta} \sqrt{det(I(\theta'))}\, d\theta'}, \tag{6.15}$$

nous pouvons facilement constater que (6.15) coincide exactement avec (6.12) : pour des familles exponentielles, *pour n grand*, l'approche Bayesienne avec une distribution *a priori* de Jeffrey est équivalente au principe du MDL.

Quand les modèles $\mathcal{H}_i$ n'appartiennent pas à la famille exponentielle, l'expression suivante est valable sous des conditions de régularité :

$$\begin{aligned} -\log \overline{p}(x^{(n)}|\mathcal{H}_i) &= -\log p(x^{(n)}|\hat{\theta}^i(x^{(n)})) + \frac{k}{2}\log\frac{n}{2\pi} \\ &\quad - \log w(|\hat{\theta}^i(x^{(n)}) + \log\sqrt{\det(\hat{I}(x^{(n)})} + o(1), \end{aligned} \tag{6.16}$$

où $\hat{I}(x^{(n)})$ est l'information *observée*.

Pour les familles exponentielles, l'information observée coincide avec l'information de Fisher, et nous sommes conduits à l'expression précédante. En dehors des modèles exponentiels, si la vraie distribution des données appartient à un des modèles $\mathcal{H}_i$, l'information observée converge encore vers l'information de Fisher, et l'approche Bayesienne est encore asymptotiquement optimale.

6.6.5 Interprétation prédictive

Soit p une distribution en $\mathcal{X}^n$. Alors

$$p(x^{(n)}) = \prod_{i=1}^{n} \frac{p(x_1^i)}{p(x_1^{i-1})} = \prod_{i=1}^{n} p(x_i|x_1^{i-1}), \tag{6.17}$$

et donc

$$-\log p(x^{(n)}) = -\sum_{i=1}^{n} \log p(x_i|x_1^{i-1}). \tag{6.18}$$

Nous pouvons interpréter le terme $-\log p(x_i|x_1^{i-1})$ comme la pénalité associée à l'observation $X_i = x_i$ quand nous essayons de prédire la valeur de X_i avec la distribution

$p(\cdots|x_1^{i-1})$ construite avec les observations précédentes. Cette pénalité sera d'autant plus petite que la valeur observée aura une probabilité élevée pour ce modèle conditionnel. L'expression (6.18) nous dit donc que la longueur des mots du code associé à la distribution p est la somme des pénalités pour la prédiction de chaque valeur X_i de la séquence basée sur toutes les valeurs précédentes x_1^{i-1} (valeurs observées).

L'équation (6.17) établit une relation entre les *modèles de probabilité* pour des séquences $x^{(n)}$ et des *stratégies de prédiction*, qui associent à chaque possible séquence passée x_1^{i-1} une loi de probabilité pour la valeur future X_i. De la même façon, cette équation nous permet d'associer une loi de probabilité définie en $\mathcal{X}^n$ à des statégies de prédiction.

Soit maintenant $\mathcal{H}$ un modèle paramétrique, et $\overline{P}$ un *code universel* par rapport au modèle $\mathcal{H}$. Dans le cadre de l'estimation Bayesienne, et pour des observation i.i.d., il est bien connu que la distribution prédictive converge (quand $n \to \infty$), vers la distribution du Maximum de Vraisemblance $p(\cdot|\hat{\theta}_1^{i-1})$. Le même comportement est obtenu pour tous le codes universels, de façon que nous pouvons approximer les distributions conditionnelles $\overline{P}(\cdot|x_1^{i-1})$ par $p(\cdot|\hat{\theta}_1^{i-1})$:

$$\overline{P}(X_i|x^{i-1} - 1) \simeq p(\cdot|\hat{\theta}_1^{i-1}), \tag{6.19}$$

et donc

$$-\log \overline{P}(x^{(n)}) \simeq -\sum_{i=1}^{n} \log p(\cdot|\hat{\theta}_1^{i-1}). \tag{6.20}$$

Sous des conditions de régularité, il peut être démontré que le membre droit de cette équation peut être écrit comme

$$-\sum_{i=1}^{n} \log p(\cdot|\hat{\theta}_1^{i-1}) = -\log p\left(\cdot|\hat{\theta}(x^{(n)})\right) + \frac{k}{2}\log n + o(1),$$

qui différe par une constante de l'expression (6.12), démontrant que cette appproche prédictive conduit asymptotiquement au même choix que le principe du MDL. Cette analyse nous fourni une interprétation alternative du MDL comme choisisant le modèle pour lequel l'erreur de prédiction accumulée est minimale.

Nous remarquons finalement que l'estimée $\hat{\theta}$ en (6.20) peut ne pas être l'estimée du Maximum de Vraissemblance, pouvant être remplacée par un autre estimateur qui converge vers l'estimateur MV. Dans certains cas, comme le montre l'exemple suivant, le comportement peut même être supérieur.

Exemple 8 *Modèle de Bernoulli*
Condiérons le modèle de Bernoulli, décrit dans la page 103, où $\theta \in [0, 1]$ est la probabilité d'observer un 1. Admettons que

$$x^{(n)} = 0\,0\,1\ \cdots,$$

et donc l'estimateur du Maximum de Vraisemblance de θ est

$$\hat{\theta}_{MV}(x_1^2) = 0,$$

impliquant que

$$p(x_3|\hat{\theta}_{MV}(x_1^2)) = 0 \Rightarrow -\log p(x_3|\hat{\theta}_{MV}(x_1^2)) = -\log p(1|\hat{\theta}_{MV}(x_1^2)) = \infty,$$

démontrant que ce modèle n'est pas universel. Cependant, la distribution prédictive basée sur l'estimateur modifié suivant (proposé par Laplace)

$$\hat{\theta}_L(x_1^2) = \frac{n_1 + \lambda}{n + 2\lambda},$$

conduit à un code universel. En effet, il peut être démontré que pour $\lambda = 1$, la distribution prédictive avec cet estimateur coincide avec la distribution marginale de Bayes relativement à une distribution *a priori* uniforme pour θ. Pour $\lambda = 2$, nous obtenons l'estimateur Baysien pour la distribution *a priori* de Jeffrey pour le modèle de Bernoulli, qui atteint (asymptotiquement) le même comportement que le code universel optimal (maximum de vraisemblance normalisé).

Cette approche a des ressemblances avec la méthode du Maximum de Vraisemblance Généalisé, décrite dans la section 6.6.1. La différence fondamentale entre les deux approches, et qui explique que l'approche prédictive qui vient d'être décrite ne souffre pas des problèmes de *overfitting* associés aux tests de Vraisemblance Généralisée, est le fait qu'ici chaque observation est prédicte (codée) avec le modèle déterminé par les observations **passées**, ce qui n'est pas le cas pour le test Maximum de Vraisemblance Généralisé. Celui-ci utilise dans un premier pas *toutes les observations* pour estimer le paramètre θ, et utilise ensuite cette estimée pour coder $x^{(n)}$.

6.7 MDL Général pour la sélection de modèles paramétriques

Le matériel présenté dans le Chapître précédent considère le cas où le nombre de modèles est fini, et la solution avancée (la distribution de Shtarkov) requiert que la compléxité paramétrique des modèles soit finie. Dans ce Chapître nous présentons une version plus générale du MDL qui peut traiter des modèles avec une compléxité infinie.

6.7.1 Complexité paramétrique infinie

La complexité paramétrique des modèles les plus comuns est infinie. Un exemple important est celui des modèles Gaussiens, comme le montre l'exemple suivant.

Exemple 9 *Complexité du modèle Gaussien*
Soit $\mathcal{H}$ la famille de distributions Gaussiennes avec *variance* σ^2 *fixée* :

$$\mathcal{H}_{\sigma^2} = \left\{ p_{\sigma^2}(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ : \ \mu \in \Re \right\},$$

étendue à $x^{(n)} \in \Re^n$ par hypothèse d'indépendance (produit des densités marginales). L'estimateur du Maximum de Vraisemblance de la moyenne μ est

$$\hat{\mu}(x_1^i) = \frac{1}{i}\sum_{j=1}^{i} x_j.$$

Alors,

$$\mathcal{C}(\mathcal{H}) = \log \int_{x^{(n)} \in \Re^+} p_{\sigma^2}(x^{(n)}|\hat{\mu}(x^{(n)})\, dx^{(n)} = \infty,$$

et donc le code universel optimal n'est pas défini pour ce modèle.
La matrice de Fisher pour ce modèle est

$$I(\mu) = \frac{1}{\sigma^2},$$

et donc nous avons également

$$\int_{\Re} \sqrt{\det(I(\mu))}\, d\mu = \infty,$$

et donc le mélange Bayesien pour la distribution *a priori* de Jeffrey n'est pas défini non plus.

Cependant, si nous considérons que $\mu \in [a, b]$, avec $a, b < \infty$, la complexité pour ce modèle "limité" est finie :

$$\int_{x^{(n)}\ :\hat{\mu}(x^{(n)})\in[a,b]} p_{\sigma^2}(x^{(n)}|\hat{\mu}(x^{(n)})\, dx^{(n)} = \frac{b-a}{\sqrt{2\pi}\sigma}\sqrt{n}.$$

Soient alors les modèles emboités suivants :

$$\mathcal{H}_K = \left\{ p_{\sigma^2,K}(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ |\mu| \leq K \right\}, K \geq 0, \tag{6.21}$$

de façon que

$$\mathcal{H} = \cup_{K\geq 0} \mathcal{H}_K.$$

Maintenant, pour chaque K, $\mathcal{C}(\mathcal{H}_K) < \infty$, et donc les codes universels optimaux correspondants, $\overline{P}_{mvn}(x^{(n)}|\mathcal{H}_K)$ existent.
Nous pouvons maintenant coder les données avec un *code en deux parties*, $\overline{P}_{meta}(x^{(n)}|\mathcal{H})$, qui code dans un premier temps la constante K, et utilise ensuite le code optimal correspondant pout coder les données. Ce code a une longueur

$$-\log \overline{P}_{meta}(x^{(n)}|\mathcal{H}) = \min_K \left\{ -\log \overline{P}_{mvn}(x^{(n)}|\mathcal{H}_K) + L(K) \right\}, \tag{6.22}$$

où $L(K)$ est la longueur du mot de code utilisé pour K.

Cette approche, basée sur un codage en deux parties, est sous-optimale. La cause de la non-optimalité est liée au fait que le code réserve plusieurs mots de code pour

la même séquence, une pour chaque valeur possible de K. Une alternative à cette approche, qui est basée dans la restriction de l'espace d'observations, consiste à limiter l'espace des paramètres. Dans [7] Rissanen propose une approche alternative, basée sur l'utilisation d'une version re-normalisée du code universel optimal :

$$\overline{P}_{rmvn}(x^{(n)})|\mathcal{H}) = \frac{\overline{P}_{mvn}\left(x^{(n)}|\mathcal{H}_{|\hat{\mu}(x^{(n)})}\right)}{\int_{y^{(n)}\in\mathcal{X}^n}\overline{P}_{mvn}\left(y^{(n)}|\mathcal{H}_{|\hat{\mu}(x^{(n)})}\right)\ dy^{(n)}}$$

6.7.2 Sommaire

Le matériel présenté indique que si nous souhaitons appliquer le principe du MDL pour choisir entre différents modèles, nous devons chercher à définir un modèle universel pour l'ensemble de modèles, qui soit capable de coder toutes les séquences d'une taille donnée n.

Si l'ensemble de modèles est fini, nous utilisons une distributions *a priori* **uniforme** pour les modèles (longueur de code constante). Dans le cas contraire, la distribution uniforme n'existe plus, et nous sommes forcés à donner une préférence à certains modèles sur les autres.

Quand la complexité paramétrique des modèles est infinie, et donc le modèle universel optimal de Shtarkov n'existe pas, nous devons décomposer $\mathcal{H}$ comme l'union de sous-modèles emboités $\mathcal{H}_k$, de complexité croissante en k. Un code universel pour l'ensemble de modèles est ensuite construit, avec une pénalité qui est proche de la pénalité associée au sous-modèle de complexité minimale qui contient l'estimée du Maximum de Vraissemblance.

Bibliographie

[1] Vijay Balasubramanian, “A geometric framework for Occam’s razor for inference of parametric distributions,” Princeton Physics Preprint PUPT-1588, Princeton, NJ, USA, 1996. (http ://arxiv.org/pdf/adap-org/9601001).

[2] J. Rissanen, “Modelling by shortest data description,” *Automatica*, 14 :465 :471, 1978.

[3] A. Kolmogorov, “Three approaches to the quantitative definition of information,” *Problems Inform. Transmission*,” (1), 1 :7, 1965.

[4] Ray Solomonoff, "A Formal Theory of Inductive Inference, Part I (II),” *Information and Control*, Part I : Vol 7, No. 1(2), pp. 1 :22 (224 :254), March(June) 1964. (http ://world.std.com/ rjs/1964pt1.pdf/http ://world.std.com/ rjs/1964pt2.pdf)

[5] Gregory Chaitin, “On the length of programs for computing finite binary sequences,” *Journal of the ACM* 13 (1966), pp. 547-569. (http ://www.cs.auckland.ac.nz/CDMTCS/chaitin/acm66.pdf)

[6] Paul Vitanyi, Ming Li, “Minimum description Length Induction, Bayesianism and Kolmogorov Complexity,” IEEE Trans. Inf. the., 46 :2, 446 :464, 2000.(http ://www.cwi.nl/ paulv/papers/mdlindbayeskolmcompl.pdf)

[7] J. Rissanen, “Strong Optimality of the Normalized ML Models as Universal Codes and Information in Data,” IEEE Trans. Information Theory, vol. 47(5), pp. 1712-1717, 2001. (http ://www.mdl-research.org/pub/bound2.ps)