UN PREMIER PAS POUR CRAQUER UN CODE SECRET

1. Introduction

Dans le film « The imitation game » l'équipe de Alan Turing parvient à faire fonctionner sa machine à décrypter les messages secrets de l'armée allemande grâce à une indication décisive qui lui permet d'identifier certaines chaînes de caractères et ainsi de guider l'algorithme pour trouver la clef de codage avant que celle-ci ne soit modifiée. Cette scène illustre que des indications peuvent permettre de casser un code de manière plus efficace ; des techniques d'algèbre linéaire peuvent aider à détecter ces indications utiles. Bien sûr les clefs de codage sont plus sophistiquées, mais on va se limiter ici à la situation d'une simple permutation des lettres de l'alphabet. Par exemple, si on se contente d'inverser l'alphabet, le texte

(1) *je suis fantastique*

devient

(2) *qv hfrh uzmgzhgrjfv.*

Afin d'analyser un texte, on lui associe la matrice $A \in \mathcal{M}_n$, où n est le nombre de lettres de l'alphabet, de ses digrammes : A_{ij} donne le nombre de fois où la lettre numéro i est suivie de la lettre numéro j. Par la suite, on construira cette matrice en supposant de plus que la dernière lettre du texte est suivie par la première (dans l'exemple *je suis fantastique* le e final est suivi du j initial). On note $e = (1, 1, ..., 1) \in \mathbb{R}^n$. Alors les produits Ae et $A^\intercal e$ donnent tous deux le vecteur f du nombre d'apparition de chaque lettre de l'alphabet. On peut comparer les fréquences $\frac{f}{f^\intercal e}$ du texte (crypté) considéré avec les fréquences observées dans des textes (non cryptés) de référence pour obtenir de premières indications utiles.

On peut aller plus loin en exploitant la décomposition en valeurs singulières de la matrice des digrammes $A = U\Sigma V^\intercal$. Si $A = \sigma u v^\intercal$ est une matrice de rang un, alors on obtient

$$Ae = \sigma(v^\intercal e)u = f = A^\intercal e = \sigma(u^\intercal e)v. \tag{3}$$

Autrement dit, dans ce cas particulier, les vecteurs u, v et f sont colinéaires. La figure 1 met en évidence cette propriété pour un texte donné : bien que la matrice des digrammes correspondantes ne soit pas de rang un, on observe une indéniable proportionnalité entre ces vecteurs !

2. Voyelles et consonnes

L'alphabet est décomposé en voyelles et consonnes qui définissent deux vecteurs v et c tels que $e = v + c$, $v^\intercal c = 0$: $v_i = 1$ et $c_i = 0$ (resp. $c_i = 1$ et $v_i = 0$) lorsque la lettre indicée par i est une voyelle (resp. consonne). À l'aide de la matrice A, on peut calculer :

- le nombre de voyelles utilisées $v^\intercal Ae = v^\intercal A(v + c)$,
- le nombre de consonnes utilisées $c^\intercal Ae = c^\intercal A(v + c)$,
- le nombre de fois où une voyelle est suivie d'une voyelle $v^\intercal Av$,
- le nombre de fois où une consonne est suivi d'une voyelle $c^\intercal Av$,...

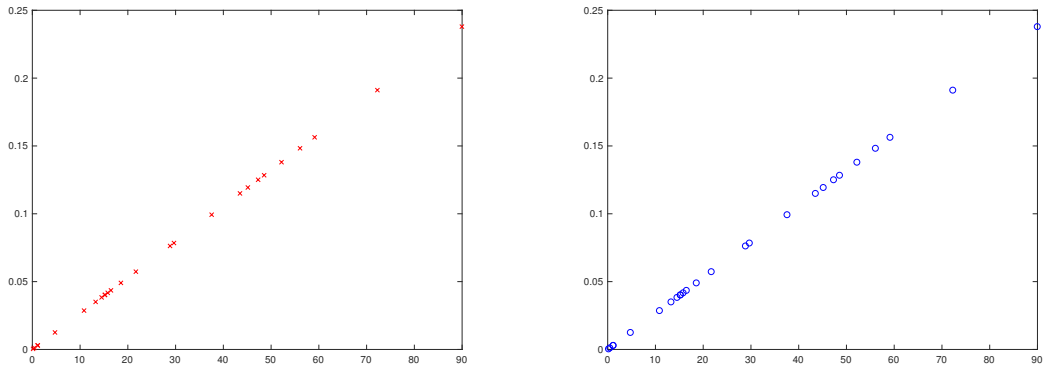

FIGURE 1. Graphe de u et de v en fonction de f pour les premières pages de *1984 de Georges Orwell*

Reconnaître cette décomposition de l'alphabet dans un texte crypté fournit de nouvelles indications utiles pour casser le code. Dans de nombreux langages s'applique la règle *vfc* selon laquelle il est plus fréquent qu'une consonne soit suivie d'une voyelle qu'une voyelle soit suivie d'une voyelle, ce qui s'exprime par la relation

$$\frac{\text{nombre de paires voyelle-voyelle}}{\text{nombre de voyelles}} \ll \frac{\text{nombre de paires consonne-voyelle}}{\text{nombre de consonnes}}. \tag{4}$$

Le hawaïen est une langue strictement *vfc* ; l'anglais est à prédominance *vfc*. Certaines caractéristiques sont plus ou moins typiques suivant les langues : par exemple, en anglais les lettres l, n, m et r sont souvent suivis de consonnes et les combinaisons *ch*, *gh*, *ph*, *sh* ou plus encore *th* sont fréquentes. En termes matriciels, la règle *vfc* s'exprime

$$\frac{v^\intercal A v}{v^\intercal A(v+c)} \ll \frac{c^\intercal A v}{c^\intercal A(v+c)}. \tag{5}$$

Lorsqu'on est confronté à un texte crypté, on en sait pas a priori identifier les consonnes et les voyelles de ce texte. Aussi, l'idée est de chercher une décomposition de l'alphabet, c'est-à-dire des vecteurs v et c dont les coordonnées valent 0 ou 1, telle que

$$v^\intercal A v \times c^\intercal A c < v^\intercal A c \times c^\intercal A v. \tag{6}$$

À cette fin, on fait appel à la décomposition SVD de la matrice A. Pécisément, on utilise l'approximation de rang 2 :

$$A_2 = \sigma_1 u_1 v_1^\intercal + \sigma_2 u_2 v_2^\intercal \tag{7}$$

et on pose

- $c_i = 1$ si $u_{2i} > 0$ et $v_{2i} < 0$, $c_i = 0$ sinon,
- $v_i = 1$ si $u_{2i} < 0$ et $v_{2i} > 0$, $v_i = 0$ sinon,
- $n_i = 1$ si u_{2i} et v_{2i} sont de même signe, $n_i = 0$ sinon.

Dans cette décomposition le vecteur n collecte les lettres « neutres » qu'on ne parvient pas à caractériser comme « consonne » ou « voyelle » par ce critère. En effet, on a

$$\begin{aligned} & v^\intercal A_2 c \times c^\intercal A_2 v - v^\intercal A_2 v \times c^\intercal A_2 c \\ & \quad = v^\intercal(\sigma_1 v_1^\intercal c u_1 + \sigma_2 v_2^\intercal c u_2) \times c^\intercal(\sigma_1 v_1^\intercal v u_1 + \sigma_2 v_2^\intercal v u_2) \\ & \qquad - v^\intercal(\sigma_1 v_1^\intercal v u_1 + \sigma_2 v_2^\intercal v u_2) \times c^\intercal(\sigma_1 v_1^\intercal c u_1 + \sigma_2 v_2^\intercal c u_2) \\ & \quad = (\sigma_1 v_1^\intercal c u_1^\intercal v + \sigma_2 v_2^\intercal c u_2^\intercal v) \times (\sigma_1 v_1^\intercal v u_1^\intercal c + \sigma_2 v_2^\intercal v u_2^\intercal c) \\ & \qquad - (\sigma_1 v_1^\intercal v u_1^\intercal v + \sigma_2 v_2^\intercal v u_2^\intercal v) \times (\sigma_1 v_1^\intercal c u_1^\intercal c + \sigma_2 v_2^\intercal c u_2^\intercal c). \end{aligned} \tag{8}$$

Dans cette expression,

— le terme en σ_1^2 est multiplié par

$$v_1^\intercal c u_1^\intercal v v_1^\intercal v u_1^\intercal c - v_1^\intercal v u_1^\intercal v v_1^\intercal c u_1^\intercal c = 0, \tag{9}$$

— le terme en σ_2^2 est multiplié par

$$v_2^\intercal c u_2^\intercal v v_2^\intercal v u_2^\intercal c - v_2^\intercal v u_2^\intercal v v_2^\intercal c u_2^\intercal c = 0. \tag{10}$$

Il ne reste donc que le terme proportionnel au produit $\sigma_1\sigma_2$, qui fait intervenir

$$\mathcal{C} = v_1^\intercal c u_1^\intercal v v_2^\intercal v u_2^\intercal c + v_1^\intercal v u_1^\intercal c v_2^\intercal c u_2^\intercal v - v_1^\intercal v u_1^\intercal v v_2^\intercal c u_2^\intercal c - v_1^\intercal c u_1^\intercal c v_2^\intercal v u_2^\intercal v. \tag{11}$$

La partition doit assurer que cette quantité est positive. Or, la matrice A a ses coefficients positifs ou nuls, il en va donc de même pour $A^\intercal A$ et $AA^\intercal$. Aussi, en vertu du théorème de Perron-Frobenius, les composantes de v_1 et de u_1 sont positives ou nulles. Comme les composantes de v et de c valent 0 ou 1, il en résulte que les termes $v_1^\intercal c$, $u_1^\intercal v$, $v_1^\intercal v$ et $u_1^\intercal c$ sont positifs ou nuls. La définition proposée pour v et c implique que

$$v_2^\intercal v \geq 0, \quad u_2^\intercal c \geq 0, \quad v_2^\intercal c \leq 0, \quad u_2^\intercal v \leq 0, \tag{12}$$

ce qui assure finalement que $\mathcal{C} \geq 0$ [1].

On applique cette méthode sur les 3002 premiers caractères du livre *1984* de George Orwell : *It was a bright cold day in April... at any rate they could plug in your wire whenever they wanted to*. Les résultats sont collectés dans la figure 2. Quatre des voyelles sont bien identifiées, une est perçue comme une consonne, quatorze consonnes sont correctement identifiées, une est perçue comme une voyelle, cinq consonnes et une voyelle ne parviennent pas à être affectées dans une catégorie déterminée. Le fait que *h* soit neutre correspond toutefois bien à son emploi en langue anglaise. La méthode donne des résultats probants, bien que la SVD d'ordre 2 (qui peut être déterminée par exemple en exploitant l'algorithme de la puissance) ne soit certainement qu'une piètre approximation de la matrice des digrammes, voir figure 3.

3. Cryptographie

On considère un codage réalisé par une simple permutation de l'alphabet : la lettre d'indice i est remplacée par la lettre d'indice $\pi(i)$. On note Π la matrice de permutation associée qui a pour $i$ème ligne, la $\pi(i)$ème ligne de la matrice identité, soit $\Pi_{ij} = \delta_{\pi(i),j}$. En particulier, on a $\sum_{\ell=1}^n \Pi_{\ell i}\Pi_{\ell j} = \sum_{\ell=1}^n \delta_{\pi(\ell),i}\delta_{\pi(\ell),j} = \delta_{ij}$ c'est-à-dire que $\Pi^\intercal = \Pi^{-1}$. La matrice

$$B = \Pi A \Pi^\intercal, \tag{13}$$

compte les digrammes du message codé : son coefficient B_{ij} donne le nombre de fois où dans le message original la lettre d'indice $\pi^{-1}(i)$ est suivie de la lettre d'indice $\pi^{-1}(j)$. Il en résulte que $B^\intercal B$ et $A^\intercal A$ ont même polynôme caractéristique et mêmes valeurs propres, et donc A et B ont les mêmes valeurs singulières. De même on obtient une décomposition SVD de B à partir de celle de A, en appliquant la permutation Π aux vecteurs singuliers de A. Ceci assure notamment que le critère de sélection des voyelles et des consonnes s'applique autant au message original qu'au message codé : il y a équivalence entre la classification de α_i dans le message original et celle de $\alpha_{\pi(i)}$ dans le message codé.

Le texte codé fournit permet de tester l'efficacité de cette approche : la figure 2 présente les fréquences des 978 lettres du texte et la décomposition en voyelles/consonnes et neutres du message codé. En comparant avec un texte de référence, les indications ainsi obtenues

1. La partition où on inverse les définitions de v et c assure aussi cette propriété.

	FREQ
J	0.0003
X	0.0003
Q	0.0013
Z	0.0020
K	0.0077
V	0.0143
B	0.0153
P	0.0193
Y	0.0200
M	0.0210
G	0.0223
F	0.0240
C	0.0273
U	0.0276
W	0.0343
D	0.0423
L	0.0446
S	0.0533
R	0.0560
H	0.0610
N	0.0680
I	0.0690
A	0.0726
O	0.0779
T	0.0943
E	0.1239

	CONS	VOY	NEU	FREQ
A	0	1	0	0.0726
B	1	0	0	0.0153
C	1	0	0	0.0273
D	1	0	0	0.0423
E	0	1	0	0.1239
F	0	0	1	0.0240
G	1	0	0	0.0223
H	0	0	1	0.0610
I	0	1	0	0.0690
J	0	1	0	0.0003
K	1	0	0	0.0077
L	1	0	0	0.0446
M	1	0	0	0.0210
N	0	0	1	0.0680
O	0	1	0	0.0779
P	1	0	0	0.0193
Q	0	0	1	0.0013
R	1	0	0	0.0560
S	1	0	0	0.0533
T	1	0	0	0.0943
U	0	0	1	0.0276
V	1	0	0	0.0143
W	1	0	0	0.0343
X	0	0	1	0.0003
Y	1	0	0	0.0200
Z	1	0	0	0.0020

original	FREQ CODE		Voyelle	Consonne	Neutre
G	0	n	0	0	0
Q	0	q	0	0	1
J	0.0010	z	0	0	1
X	0.0020	e	0	0	1
V	0.0072	p	0	1	0
K	0.0092	a	0	0	1
B	0.0123	b	0	1	0
O	0.0143	j	0	1	0
W	0.0164	g	0	1	0
C	0.0174	w	0	1	0
M	0.0174	l	1	0	0
Y	0.0194	c	0	0	1
S	0.0297	h	0	1	0
F	0.0307	r	0	1	0
Z	0.0317	m	0	1	0
L	0.0368	s	0	1	0
R	0.0460	f	0	1	0
P	0.0542	i	0	1	0
D	0.0624	o	0	1	0
H	0.0644	u	0	0	1
U	0.0685	k	1	0	0
I	0.0736	v	1	0	0
N	0.0787	x	0	0	1
A	0.0818	d	1	0	0
T	0.0900	t	0	1	0
E	0.1350	y	1	0	0

FIGURE 2. À gauche : Analyse de *1984 de George Orwell* : tri des lettres par fréquence croissante et partition obtenue. À droite : Analyse d'un texte codé (extrait *The Man Who Would Be A King* de Rudyard Kipling).

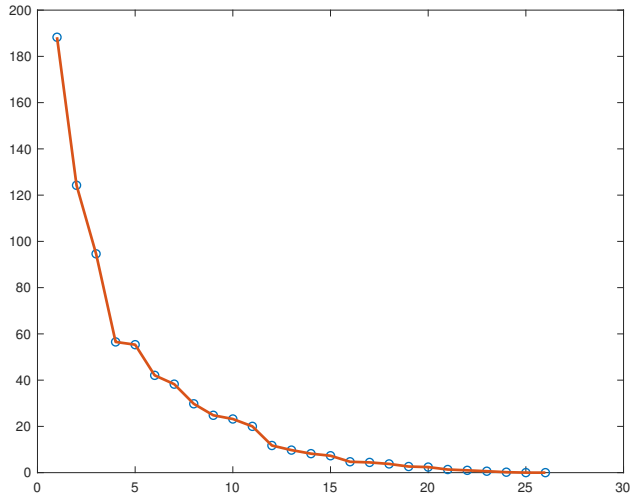

FIGURE 3. Analyse de *1984 de George Orwell* : évolution des valeurs singulières

permettent une avancée substantielle vers le décodage du texte. La figure révèle aussi la clef de codage, montrant l'efficacité de cette approche.

En pratique les codes n'utilisent pas une simple permutation mais plutôt un ensemble de K permutations $\Pi_1, ..., \Pi_K$ utilisées de manière cyclique. Ainsi, la $m^{\text{ème}}$ lettre α_i, $i \in \{1, ..., m\}$, du texte original est codée comme $\alpha_{\Pi_p(i)}$ où m est congru à p modulo K. Il existe des techniques qui permettent de deviner la valeur K du cycle, puis d'appliquer les stratégies d'identification des digrammes et de la partition en voyelles et consonnes.