Profils de vote : expérience et modélisation

François Durand, sous la direction de Jean-François Laslier

Stage de recherche de maîtrise du MMFAI
effectué de juin à août 2003
au Laboratoire d'Économétrie de l'École Polytechnique
1 rue Descartes, 75005 Paris

Table des matières

Introduction

Dans une démocratie, le choix des modes de scrutin est bien sûr primordial. Même pour un problème aussi simple que l'élection d'un chef d'État, il existe de nombreux modes de scrutin possibles, qui peuvent donner des résultats différents pour une même répartition des votes.

Comment choisir ? On s'attend à ce qu'un « bon » système de vote vérifie un certain nombre de propriétés, comme le fait de ne pas laisser toute la décision à un seul électeur (système non dictatorial). Malheureusement, dès qu'on exige certaines propriétés, apparemment naturelles, aucun système de vote ne les satisfait.

Par exemple, on peut exiger les propriétés suivantes :

- Unanimité (principe de Pareto) : si tous les électeurs préfèrent A à B, alors le mode de scrutin classe A devant B. En particulier, B ne peut pas être élu.
- Indépendance des choix non significatifs : si certains électeurs changent d'avis, mais pas en ce qui concerne la comparaison entre A et B, alors le mode de scrutin classe toujours de la même façon A et B.
- Non-dictature : le résultat du vote ne dépend pas seulement des choix d'un seul électeur !

Il n'existe pas de mode de scrutin qui vérifie ces propriétés (ou plus exactement de fonction de choix social) : c'est le théorème d'Arrow.

Certaines des propriétés présentées peuvent ne pas paraître évidentes... Mais en fait, on peut prouver un résultat encore plus frappant que le théorème d'Arrow. Exigeons seulement les propriétés suivantes :

- Non-manipulabilité : les électeurs ont toujours intérêt à exprimer leurs vraies préférences ; ne pas exprimer leur opinion réelle ne peut en aucun cas mener à un résultat qui les satisfait davantage.
- Non-dictature : le résultat ne dépend pas que d'un électeur.

Dès qu'il y a au moins trois candidats, le théorème de Gibbard-Satterthwaite énonce qu'aucun mode de scrutin ne vérifie ces propriétés. Un tel résultat est particulièrement décourageant pour établir une démocratie, puisqu'il signifie que dans tout système de vote, certains électeurs peuvent avoir intérêt à ne pas exprimer leurs préférences... À partir de là, comment le mode de scrutin peut-il faire la synthèse des avis individuels en un choix collectif ? C'est impossible, puisque que les urnes ne « connaissent » pas les avis individuels !

Ce genre de paradoxes, qu'on peut qualifier d'intrinsèques à chaque mode de scrutin, entraîne des paradoxes entre les systèmes : deux systèmes sensés peuvent donner des résultats différents, selon que le système privilégie davantage l'une ou

l'autre des propriétés ci-dessus, voire d'autres propriétés issues de considérations politiques ou culturelles.

Le choix du mode de scrutin est donc un problème épineux, tout simplement parce qu'il est insoluble. Cependant, la situation est peut-être moins désespérée qu'elle n'en a l'air...

On appelle culture les préférences d'une population entière sur un sujet donné, c'est-à-dire ce qu'on va trouver dans les urnes si on demande aux gens leur liste de préférences. Les théorèmes précédents signifient qu'il peut y avoir des paradoxes électoraux pour certaines cultures, mais ce n'est pas toujours le cas : pour prendre l'exemple le plus simple, dans la culture où tout le monde est d'accord, il est facile de trouver un système de vote juste et équitable !

En fonction des types de cultures, les problèmes posés par le mode de scrutin varient. On n'a pas forcément les mêmes paradoxes, et en tous cas pas avec la même probabilité. C'est pourquoi il est particulièrement intéressant de connaître les structures d'opinion réalistes : cela permet d'évaluer les problèmes posés par les différents modes de scrutins, et de choisir en meilleure connaissance de cause.

Pour ce faire, nous avons réalisé une expérience de vote, que nous décrivons dans le premier chapitre. Notre but est de dégager des profils de culture théoriques qui expliquent bien les résultats constatés.

Pour chaque vote, on demande à l'internaute de classer les différentes options : sa réponse est donc un élément du groupe des permutations $\mathcal{S}_n$. Dans le deuxième chapitre, nous présentons une distance sur $\mathcal{S}_n$ adaptée à l'analyse des votes.

Dans le troisième chapitre, nous présentons la méthode du maximum de vraisemblance, qui permet de comparer entre eux les différents modèles, et ses liens avec l'information et l'entropie.

Dans le quatrième chapitre, nous développons un modèle monopôle relativement simple : une permutation est privilégiée, et la probabilité décroît exponentiellement quand on s'éloigne de ce pôle.

Dans le cinquième chapitre, nous testons différentes lois d'une complexité d'écriture donnée pour comparer leurs performances respectives et enfin valider la loi exponentielle.

La crédibilité de cette hypothèse étant vérifiée, nous pouvons envisager de passer à un modèle multipolaire mais, pour limiter les problèmes de complexité algorithmique, nous nous limitons à deux pôles.

L'annexe A présente un tableau synthétique des résultats obtenus, particulièrement utile pour comparer les différents modèles entre eux. L'annexe B présente les spécifications de quelques unes des principales fonctions MatLab que nous avons écrites pour l'analyse des résultats.

Différentes méthodes ont déjà été utilisées pour analyser des structures d'opinion mais, à notre connaissance, les profils théoriques de probabilité sur $\mathcal{S}_n$ ne sont guère utilisés à l'heure actuelle, et encore moins validés par l'expérience. Cette recherche est donc un préliminaire à l'étude du comportement des modes de scrutin dans des situations réalistes.

Chapitre 1

Site web de vote online

Un site web a été créé pour recueillir des votes réels. Son adresse est :

`http://experiencedevote.free.fr`

1.1 Contenu du site

Pour chaque vote, on demande à l'internaute de classer un certain nombre d'options. Son opinion est donc représentée par un élément du groupe des permutations.

Cette structure mathématique présente les avantages suivants : elle donne une image riche de l'opinion de l'électeur, mais sans ajouter d'information parasite, comportant une part d'abitraire, indépendante des préférences de l'électeur sur le sujet donné, comme une notation.

Ci-dessous, vous trouverez la liste des questions posées. Chacune était accompagnée d'un mode d'emploi soulignant en particulier qu'il fallait ordonner toutes les options, et éventuellement d'un petit texte explicatif qui précisait la question.

1.1.1 Revotez pour les candidats de 2002.

- Alain MADELIN (DL) ;
- Arlette LAGUILLER (LO) ;
- Bruno MÉGRET (MNR) ;
- Christiane TAUBIRA (RPG) ;
- Christine BOUTIN (FRS) ;
- Corinne LEPAGE (Cap 21) ;
- Daniel GLUCKSTEIN (PT) ;
- François BAYROU (UDF) ;
- Jacques CHIRAC (RPR) ;
- Jean SAINT-JOSSE (CPNT) ;
- Jean-Marie LE PEN (FN) ;
- Jean-Pierre CHEVÈNEMENT (MDC) ;
- Lionel JOSPIN (PS) ;

– Noël MAMÈRE (Les Verts) ;
– Olivier BESANCENOT (LCR) ;
– Robert HUE (PCF).

1.1.2 Quel mode de scrutin direct vous semble le plus juste pour la présidentielle ?

– assentiment ;
– Borda ;
– Condorcet ;
– notation ;
– uninominal à un tour ;
– uninominal à deux tours.

1.1.3 Parmi les pays suivants, lesquels sont les plus développés ?

– Brésil ;
– Guatemala ;
– Haïti ;
– Lituanie ;
– Mozambique ;
– Norvège ;
– Singapour ;
– Tunisie.

1.1.4 Si vous ne pouviez faire qu'une chose « utile » de votre vie, laquelle choisiriez-vous ?

– élever un enfant ;
– empêcher l'oubli d'une connaissance scientifique ;
– empêcher la démolition d'un monument historique ;
– sauver des flammes un tableau de maître ;
– sauver une vie humaine.

1.1.5 Si vous n'emportiez qu'une cassette dans une île déserte, laquelle choisiriez-vous ?

– Blanche-Neige et les sept nains ;
– Le Bon, la Brute et le Truand ;
– Casablanca ;
– Le fabuleux destin d'Amélie Poulain ;
– La guerre des étoiles ;
– La liste de Schindler ;
– Psychose ;
– Singin' in the Rain.

1.1.6 Quelle est votre couleur de chaussettes préférée ?

- blanc ;
- bleu ;
- jaune ;
- noir ;
- rouge ;
- vert.

1.2 Commentaires sur le protocole expérimental

On a en général pris garde à placer l'internaute devant une situation de choix effectif, et non de préférence. Ce point nous semble important car, si on aime différents films pour des raisons différentes, *a priori* non comparables, l'expérience de pensée de l'île déserte permet de se mettre en situation de décision, ce qui est en général le but d'un vote.

Cependant, nous ne l'avons pas toujours fait, en particulier pour le vote sur les chaussettes, et ce serait sans doute préférable pour une expérience future. En ce qui concerne les pays développés, ce n'est pas le cas non plus, mais le texte explicatif permettait de cibler la question d'une autre façon, en faisant référence à l'Indice de Développement Humain.

Le choix des votes a été difficile, car nous voulions que chaque vote porte sur un sujet à propos duquel chaque électeur ait ou puisse se faire un avis. Ceci excluait en particulier certaines questions trop « culturelles », pour lesquelles certains électeurs auraient pu ne pas connaître telle ou telle option du vote. Nous n'avons pu échapper totalement à cet écueil, en particulier pour les films, mais nous avons cherché à ne proposer que des films dont la plupart des électeurs ont une idée, même s'ils ne les ont pas vus. Ces films ont été choisis parmi les plus reconnus, à partir d'Internet Movie Database.

Nous voulions également que les votes portent sur des sujets aussi divers que possibles, sérieux ou moins sérieux, pour que la recherche soit aussi dégagée du sujet traité que possible : trouver des profils de cultures semblables pour des votes totalement différents nous paraissait plus concluant. C'est pourquoi on va de la politique au cinéma, en passant par la couleur des chaussettes.

Les réponses étaient présentées à l'internaute dans l'ordre alphabétique, et ceci peut introduire un biais dans l'expérience. Dans une version ultérieure, il serait sans doute préférable de permuter aléatoirement les options. Ceci dit, tout système de vote réel ne présente-t-il pas ce genre de biais ?

1.3 Résultats

Pour se faire une première idée des résultats du site, on peut donner les matrices de vote. Le chiffre m_{ij} sur la ligne i et la colonne j est le nombre de personnes ayant placé i devant j dans leur choix. La colonne V représente le nombre de victoires de chaque candidat dans le tournoi majoritaire. La colonne B est le score de Borda de chaque candidat, c'est-à-dire la somme sur j des m_{ij}. Pour améliorer la lisibilité, nous avons ordonné les options dans l'ordre médian,

terme que nous définirons par la suite.

1.3.1 Vote sur la présidentielle

Il y a eu 60 votants.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	V	B
1. JOSPIN	0	42	46	46	49	47	44	50	50	49	52	53	55	56	58	58	15	755
2. TAUBIRA	18	0	31	41	36	38	41	43	44	48	44	52	53	54	58	58	14	659
3. MAMÈRE	14	29	0	44	38	37	44	44	43	47	49	51	51	52	57	56	13	656
4. HUE	14	19	16	0	31	30	35	35	38	44	42	53	47	48	56	55	11	563
5. BAYROU	11	24	22	29	0	33	34	40	43	44	52	47	58	57	58	58	11	610
6. CHEVÈNEMENT	13	22	23	30	27	0	34	38	41	42	47	46	55	54	58	58	10	588
7. BESANCENOT	16	19	16	25	26	26	0	30	31	49	38	49	44	40	57	56	8	522
8. LEPAGE	10	17	16	25	20	22	30	0	36	40	43	45	55	55	57	57	8	528
9. CHIRAC	10	16	17	22	17	19	29	24	0	38	48	40	52	50	58	58	7	498
10. LAGUILLER	11	12	13	16	16	18	11	20	22	0	32	38	40	38	56	55	6	398
11. MADELIN	8	16	11	18	8	13	22	17	12	28	0	32	35	37	56	58	5	371
12. GLUCKSTEIN	7	8	9	7	13	14	11	15	20	22	28	0	36	38	55	54	4	337
13. BOUTIN	5	7	9	13	2	5	16	5	8	20	25	24	0	38	56	56	3	289
14. SAINT-JOSSE	4	6	8	12	3	6	20	5	10	22	23	22	22	0	57	57	2	277
15. LE PEN	2	2	3	4	2	2	3	3	2	4	4	5	4	3	0	33	1	76
16. MÉGRET	2	2	4	5	2	2	4	3	2	5	2	6	4	3	27	0	0	73

1.3.2 Vote sur les modes de scrutin

Il y a eu 34 votants.

	1	2	3	4	5	6	V	B
1. Condorcet	0	25	17	21	28	28	4	119
2. Borda	9	0	17	17	27	29	2	99
3. Assentiment	17	17	0	20	20	24	3	98
4. Notation	13	17	14	0	24	25	2	93
5. Uninominal à deux tours	6	7	14	10	0	27	1	64
6. Uninominal à un tour	6	5	10	9	7	0	0	37

1.3.3 Vote sur les pays développés

Il y a eu 49 votants.

	1	2	3	4	5	6	7	8	V	B
1. Norvège	0	45	49	49	49	49	49	49	7	339
2. Singapour	4	0	39	42	45	49	48	48	6	275
3. Lituanie	0	10	0	26	30	47	47	46	5	206
4. Brésil	0	7	23	0	28	47	48	48	4	201
5. Tunisie	0	4	19	21	0	48	49	48	3	189
6. Guatemala	0	0	2	2	1	0	41	39	2	85
7. Mozambique	0	1	2	1	0	8	0	25	1	37
8. Haïti	0	1	3	1	1	10	24	0	0	40

1.3.4 Vote sur les actions utiles

Il y a eu 56 votants.

	1	2	3	4	5	V	B
1. Sauver une vie humaine	0	29	29	41	40	4	139
2. Empêcher l'oubli d'une connaissance scientifique	27	0	31	47	44	3	149
3. Élever un enfant	27	25	0	38	37	2	127
4. Empêcher la démolition d'un monument historique	15	9	18	0	39	1	81
5. Sauver des flammes un tableau de maître	16	12	19	17	0	0	64

1.3.5 Vote sur les films

Il y a eu 49 votants.

	1	2	3	4	5	6	7	8	V	B
1. Le fabuleux destin d'Amélie Poulain	0	26	31	30	27	34	37	37	7	222
2. La guerre des étoiles	23	0	26	24	30	34	34	31	5	202
3. Le Bon, la Brute et le Truand	18	23	0	29	27	29	36	34	5	196
4. Casablanca	19	25	20	0	25	26	26	32	5	173
5. Psychose	22	19	22	24	0	26	34	30	3	177
6. La liste de Schindler	15	15	20	23	23	0	33	30	2	159
7. Blanche-Neige et les sept nains	12	15	13	23	15	16	0	25	1	119
8. Singin' in the Rain	12	18	15	17	19	19	24	0	0	124

1.3.6 Vote sur la couleur des chaussettes

Il y a eu 53 votants.

	1	2	3	4	5	6	V	B
1. Noir	0	30	40	46	46	44	5	206
2. Blanc	23	0	36	46	50	47	4	202
3. Bleu	13	17	0	45	47	46	3	168
4. Rouge	7	7	8	0	30	33	2	85
5. Vert	7	3	6	23	0	29	1	68
6. Jaune	9	6	7	20	24	0	0	66

1.4 Considérations techniques

Le site est interactif ; deux langages de programmation ont permis de mettre en place cette interactivité. Côté serveur, c'est PHP, couplé avec Sql, qui permet d'enregistrer les votes et de générer les fichiers de données ASCII qui seront facilement utilisables par n'importe quel logiciel de traitement, par exemple Matlab. Par ailleurs, la page est créée à partir des champs de la base Sql, ce qui permet d'ajouter très facilement un nouveau vote, en créant une nouvelle table Sql.

Côté client, c'est en javascript que sont programmées les petites routines qui permettent de classer les options du vote.

Chapitre 2

Quelques considérations sur le groupe des permutations

Les choix d'un électeur sont représentés par un élément du groupe $\mathcal{S}_n$ des permutations, où n est le nombre d'options proposées. On notera $[\sigma(1)\ \ \sigma(2)\ldots\ \ \sigma(n)]$ la permutation σ. Les cycles seront notés suivant l'usage, avec des parenthèses.

2.1 Distance choisie sur $\mathcal{S}_n$

On peut définir nombre de distances sur $\mathcal{S}_n$. Pour respecter la symétrie inhérente à ce groupe, on choisit en général une distance d telle que $d(\sigma_1, \sigma_2) = d(\sigma_1 \circ \sigma_0, \sigma_2 \circ \sigma_0)$. En particulier :

$$d(\sigma_1, \sigma_2) = d(\sigma_1 \circ \sigma_2^{-1}, Id).$$

Si d est la distance de Cayley, $d(\sigma, Id)$ est par définition le nombre minimum de transpositions intervenant dans la décomposition de σ. Mais alors on a

$$d([n\ \ 2\ \ 3\ \ldots\ \ n-1\ \ 1], Id) = 1 = d([2\ \ 1\ \ 3\ \ 4\ \ \ldots\ \ n], Id).$$

Or si la première permutation échange la première et la dernière préférence, ce qui est un changement d'opinion important, la seconde n'en échange que deux consécutives, ce qui représente un changement d'opinion moins radical.

La règle de Spearmann est définie par $d(\sigma, Id) = \sum |\sigma(i) - i|$. Mais alors

$$d([n\ \ 2\ \ 3\ \ldots\ \ n-1\ \ 1], Id) = 2(n-1) = d([2\ \ 3\ \ \ldots\ \ n\ \ 1], Id).$$

Le premier changement d'opinion place le dernier en tête et le premier en queue, alors que le second changement d'opinion ne modifie que la place du premier par rapport aux autres, et laisses inchangées les places respectives de ceux-ci.

Ces considérations nous amènent à choisir une distance qui se base sur les places respectives des éléments entre eux. Plus précisément, on prend la distance de la différence symétrique ou distance de Kemeny :

$$\boxed{d(\sigma, Id) = \sum_{i<j} 1_{\sigma(i)>\sigma(j)}.}$$

Cette distance est donc définie comme la somme sur-diagonale de la matrice de vote correspondant à un seul électeur ayant voté σ. Il est facile de montrer que cette distance correspond au nombre minimal de transpositions $(i \quad i+1)$ intervenant dans la décomposition de σ.

2.2 Topologie de $\mathcal{S}_n$

Muni de cette distance, $\mathcal{S}_n$ peut être représenté par un graphe : il sufffit de placer ses éléments sur les sommets et de relier entre elles celles qui sont à distance 1. Vue la définition de la distance, la distance entre deux sommets est la longueur du plus court chemin de l'un à l'autre.

De cette façon, $\mathcal{S}_2$ sera représenté par un segment, $\mathcal{S}_3$ par un hexagone, et $\mathcal{S}_4$ par un octaèdre tronqué.

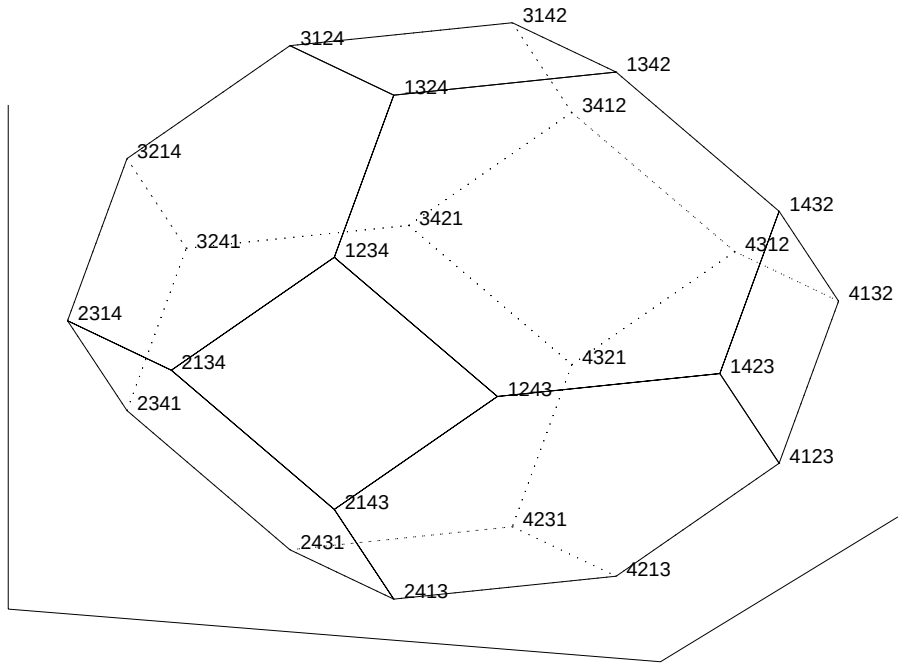

FIG. 2.1: Le groupe $\mathcal{S}_4$ des permutations de quatre éléments, muni de la distance de Kemeny d.

De manière générale, $\mathcal{S}_n$ peut être représenté par un polytope semi-régulier plongé dans l'espace à $n-1$ dimensions, dit permutoèdre.

En ce qui nous concerne, cette correspondance géométrique pourra servir à visualiser les profils des votes, avec par exemple l'apparition de « bosses » sur le polyèdre semi-régulier. C'est pourquoi, pour cette représentation, nous nous

limiterons à $\mathcal{S}_4$, qui donne des graphiques encore facilement lisibles par un œil humain.

2.3 Répartition de $d(\sigma, Id)$: présentation des T_n^k

Nous aurons besoin, par la suite, de savoir combien d'éléments de $\mathcal{S}_n$ se trouvent à une distance k d'une permutation donnée, par exemple l'identité. Notons ce nombre T_n^k.

Pour choisir une permutation σ, on peut commencer par choisir son dernier élément $\sigma(n)$. On prend un élément i et on le fait redescendre en dernière position. Ceci occasionne $n-i$ transpositions $(j \;\; j+1)$. Puis on choisit une permutation de $\mathcal{S}_{n-1}$, qu'on applique aux $n-1$ premiers éléments. Formellement, on a la relation de récurrence suivante :

$$T_n^k = \sum_{l=\max(k-n+1,0)}^{k} T_{n-1}^l.$$

En faisant la différence avec T_n^{k+1}, on trouve une relation plus simple :

$$T_n^{k+1} = T_n^k + T_{n-1}^{k+1} - T_{n-1}^{k-n+1}.$$

Cette relation permet un calcul numérique des valeurs de T_n^k souhaitées. À titre indicatif, voici les valeurs pour $n \leq 6$:

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$n=1$	1															
$n=2$	1	1														
$n=3$	1	2	2	1												
$n=4$	1	3	5	6	5	3	1									
$n=5$	1	4	9	15	20	22	20	15	9	4	1					
$n=6$	1	5	14	29	49	71	90	101	101	90	71	49	29	14	5	1

Chapitre 3

Méthode du maximum de vraisemblance

Nous allons tester plusieurs modèles pour expliquer, ou au moins décrire, les profils expérimentaux. Pour les comparer, nous allons utiliser la méthode du maximum de vraisemblance.

3.1 Principe de la méthode

Soit $\mu(\sigma)$ une loi de probabilité théorique sur $\mathcal{S}_n$. Soit $\Sigma = (\sigma_1, \ldots, \sigma_s)$ une suite de votes, par exemple le résultat d'une expérience.

En supposant que les votes sont indépendants et que le modèle théorique est vrai, la probabilité d'obtenir la suite de votes Σ vaut :

$$\mu(\Sigma) = \prod_{i=1}^{s} \mu(\sigma_i)$$

Notre but est de maximiser cette grandeur : le modèle est alors celui qui explique le mieux l'expérience, puisqu'il la rend la plus probable possible.

3.2 Lien avec l'information et l'entropie

Si nous prenons le logarithme en base 10 de l'expression précédente, nous obtenons :

$$\log(\mu(\Sigma)) = \sum_{i=1}^{s} \log(\mu(\sigma_i))$$

La distribution qui maximise cette grandeur est celle où $\mu(\sigma)$ est le nombre n_σ de votes σ divisé par s : c'est un résultat classique que l'on prouve par le théorème des extrema liés (méthode des multiplicateurs de Lagrange). La grandeur optimisée vaut alors :

$$S_{\max} = \log(\mu(\Sigma)) = \sum_{\sigma \in \mathcal{S}_n} n_\sigma \log(\frac{n_\sigma}{s})$$

Ce maximum est l'entropie du vote.

Le pire modèle que nous puissions raisonnablement envisager, c'est celui où nous ne disposons d'aucune information sur le vote : à ce moment-là, il est naturel de prendre un profil équiprobable. Le logarithme de la probabilité vaut alors :

$$S_{\min} = \log(\mu(\Sigma)) = -s \log(n!)$$

Cette borne inférieure servira de référence aux modèles que nous allons développer.

La différence entre $S_{\min}$ et $S_{\max}$ est la quantité d'information qui distingue l'expérience d'un profil uniforme. Pour évaluer de façon absolue la validité d'un modèle, nous utiliserons la grandeur suivante :

$$x = \frac{\log(\mu(\Sigma)) - S_{\min}}{S_{\max} - S_{\min}},$$

que nous exprimerons sous forme de pourcentage. C'est la part de la structure du vote qui est décrite par le modèle.

Chapitre 4

Modèle monopôle en loi exponentielle

Quand on visualise le graphe d'un vote, on constate souvent la présence de « bosses » localisées autour d'une certaine permutation. Par exemple, si on prend le vote sur la couleur des chaussettes et que, pour des raisons de visualisation, on se limite aux options {1 Blanc, 2 Bleu, 3 Jaune, 4 Noir}, on obtient le graphe suivant.

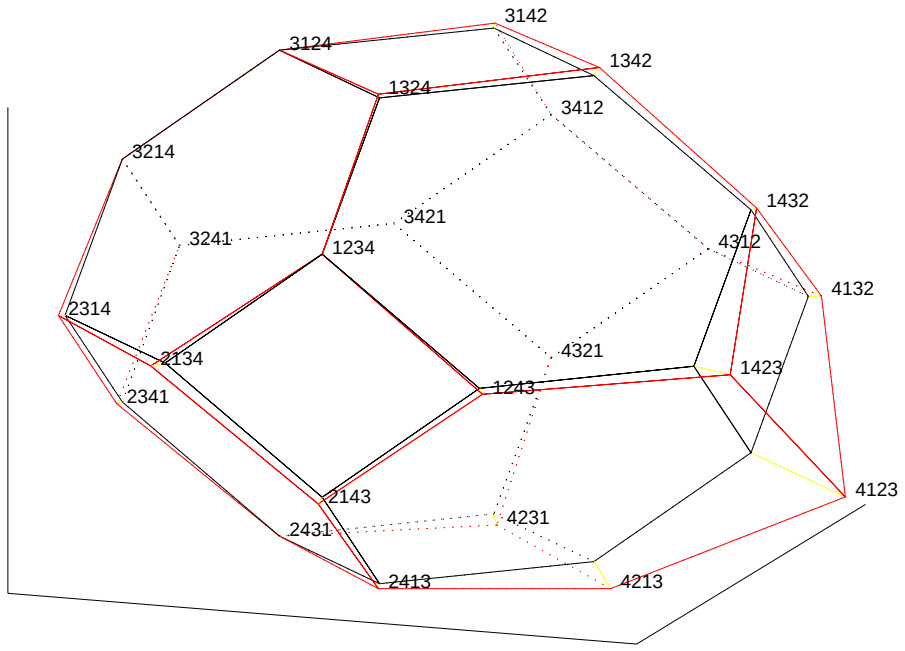

FIG. 4.1: Vote sur les chaussettes, où 1 = Blanc, 2 = Bleu, 3 = Jaune, 4 = Noir

Ce graphe se lit de la façon suivante : la distance (dans l'espace) entre le sommet du polyèdre déformé et le sommet de référence correspondant est proportionnel au nombre de votes pour le sommet.

On voit qu'il y a une permutation privilégiée, [4 1 2 3], et que le polyèdre est également déformé aux alentours de ce pôle. C'est ce que nous allons essayer de modéliser.

4.1 Présentation du modèle

On suppose que chaque électeur vote selon une même loi aléatoire et de façon indépendante. Dans ce premier modèle, il y a une « doxa », c'est-à-dire une permutation pôle σ_0 qui présente le maximum de probabilité.

On suppose que la loi de probabilité μ est de la forme :

$$\mu(\sigma) = p^{d(\sigma,\sigma_0)}/Z$$

où $Z(n,p)$ est une constante de normalisation, et $p \in [0,1]$ est le paramètre du modèle. C'est en quelque sorte une « probabilité d'erreur » par rapport à la doxa.

Le sens ou la pertinence par rapport aux données du choix d'une loi exponentielle n'est pas évident à justifier *a priori* : ceci sera développé dans le chapitre

suivant. Pour le moment, contentons-nous de constater que ce modèle offre un avantage pratique indéniable, puisqu'il permet de séparer l'optimisation de σ_0 de celle du paramètre p.

4.2 Optimisation du modèle : décomposition du problème

Dans l'optique du maximum de vraisemblance, on cherche σ_0 et p tels que, sous l'hypothèse du modèle, la probabilité des données expérimentales soit maximale.

Notons $\Sigma = (\sigma_1, \ldots \sigma_s)$ le s-uplet des votes. La probabilité de Σ est donnée par :

$$\mu(\Sigma) = p^{\sum_{i=1}^{s} d(\sigma_i, \sigma_0)} / Z^s$$

Quel que soit p, le pôle σ_0 qui optimise $\mu(\Sigma)$ est donc l'ordre médian, c'est-à-dire tel que la somme des distances aux différents votes soit minimale. On retrouve une méthode classique pour trouver un ordre optimal, celle de Condorcet-Young.

4.3 Optimisation du modèle : recherche de σ_0

On appelle matrice de vote celle dont le coefficient (i, j) est le nombre d'électeurs qui ont placé l'option i devant l'option j. On appelle matrice de vote antisymétrique M la précédente moins sa transposée.

On montre que σ_0 est la permutation des vecteurs de base qui maximise la somme des éléments sur-diagonaux de la matrice de vote antisymétrique.

S'il existe une partition des indices $G \cup P = [1 \ldots n]$ telle que pour tout $i \in G$, $j \in P$, $M(i, j) > 0$, alors on dit que G est un cycle supérieur du tournoi majoritaire. La permutation σ_0 doit alors être telle que pour tout $i \in G$, $j \in P$, $\sigma_0(i) < \sigma_0(j)$.

L'algorithme que nous avons utilisé pour déterminer σ_0 est donc le suivant :
- on cherche le plus petit cycle supérieur non vide ;
- on le classe par un algorithme glouton qui essaie toutes ses permutations ;
- on classe son complémentaire.

Dans le meilleur des cas l'algorithme est en n^2 ; dans le pire des cas, il est en $n^2.n!$.

Dans les cas pratiques que nous avons rencontrés, l'algorithme glouton seul est inutilisable, alors qu'en cherchant d'abord les cycles supérieurs, le calcul est quasi-instantané.

Dans une version ultérieure, nous pourrions aussi utiliser les options « couvertes » : si pour tout j, $M(i_1, j) > M(i_2, j)$, alors $\sigma_0(i_1) < \sigma_0(i_2)$.

4.4 Optimisation du modèle : recherche de p

Une fois σ_0 trouvé par la méthode précédente, il reste à évaluer p.

On sait calculer $\mu(\Sigma)$ en fonction de p ; il suffit donc de maximiser cette grandeur numériquement.

Dans les expériences que nous avons menées, nous avons trouvé les résultats suivants :

vote	p	$\log(\mu(\Sigma))$	x
films	$0,82$	$-218,88$	$4,75\%$
président	$0,71$	$-667,03$	$19,09\%$
mode de scrutin	$0,66$	$-88,89$	$15,90\%$
actions utiles	$0,65$	$-107,80$	$23,26\%$
couleur des chaussettes	0.50	$-121,68$	$40,81\%$
pays développés	0.33	$-125,85$	$66,05\%$

L'ordre médian σ_0 (pôle du modèle) n'a pas été redonné car c'est celui dans lequel on a placé les options dans les matrices de votes, au chapitre 1. Pour une comparaison des résultats avec les autres modèles, on pourra se reporter à l'annexe A.

Il n'est pas surprenant que la valeur la moins élevée de p soit trouvée pour le vote sur les pays développés : c'est celui pour lequel la notion de pôle a le plus de sens. Ceci dit, le pôle qui se dégage de ce vote présente une erreur par rapport au classement IDH des différents pays : les électeurs placent le Mozambique devant Haïti, alors que le Mozambique est moins développé (un des cinq IDH les plus faibles du monde).

En tous cas, le modèle donne un première description valable de la réalité, sauf peut-être pour le vote sur les films, mais ceci n'est pas étonnant non plus. C'est sans doute celui où l'on discute le moins d'opinion et le plus de préférences : les goûts et les couleurs...

Chapitre 5

Généralisation du modèle monopôle

Nous avons pour le moment utilisé, dans le cadre du modèle monopôle, une loi exponentielle décroissante. Mais nous n'avons pas justifié ce choix, sinon par les avantages pratiques qu'il offre. Nous voudrions savoir s'il est confirmé par les données expérimentales.

Nous aimerions déterminer une loi de la forme $\mu(\sigma) = f(k)$, où $k = d(\sigma, \sigma_0)$, ou bien une loi de la forme $\mu(k) = g(k)$. Les deux expressions diffèrent d'une variable multiplicative T_n^k. Pour orienter nos recherches, examinons le graphique de la répartition des votes sur les actions utiles , suivant les deux points de vue.

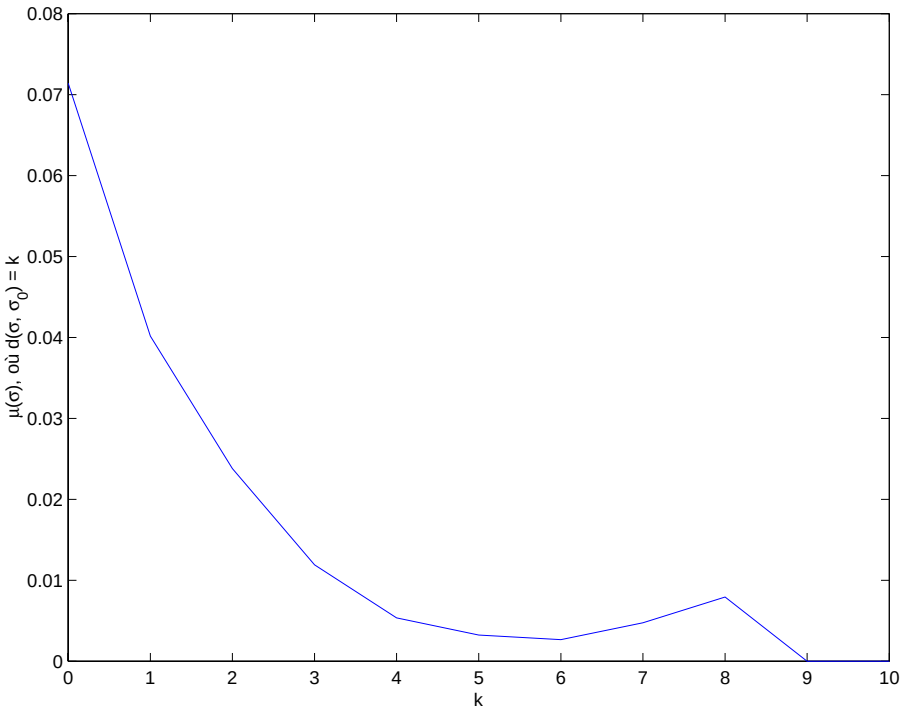

FIG. 5.1: Fréquence moyenne d'un vote donné, si celui-ci est à distance k du pôle, dans l'expérience sur les actions utiles : c'est la probabilité expérimentale d'être à distance k du pôle, divisée par T_n^k.

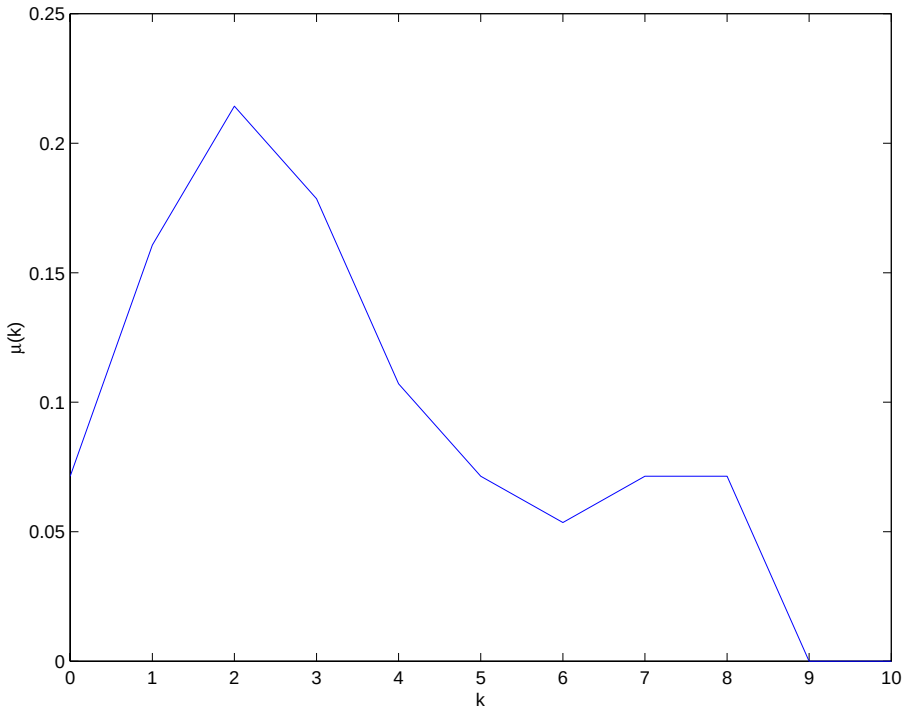

FIG. 5.2: Probabilité expérimentale d'être à distance k du pôle, pour le vote sur les actions utiles.

Le premier graphe présente une forme plus régulière : on peut donc penser qu'il sera plus facile à fitter avec un profil théorique simple, et surtout décroissant, ce qui était tout de même l'idée originale du pôle.

5.1 Aperçu de la méthode

On suppose que la probabilité suit une loi de la forme :

$$\mu(\sigma) = f(d(\sigma, \sigma_0)).$$

Pour ne pas se soucier des problèmes de normalisation, nous considérerons des lois de probabilité non normalisées, notées $F(d(\sigma, \sigma_0))$. Si n_k est le nombre de votes σ tels que $d(\sigma, \sigma_0) = k$ et s le nombre de votants, alors on a :

$$\mu(\Sigma) = \frac{\prod F(k)^{n_k}}{(\sum T_n^k F(k))^s}.$$

Il faut maximiser ce nombre.

Mais le choix possible pour F est large ! L'analyste peut proposer diverses formes de lois, des plus simples aux plus complexes, et optimiser leurs paramètres pour voir laquelle est la plus efficace, mais c'est long...

L'idée de la méthode est de faire faire ce travail à la machine elle-même. On a pensé à un algorithme génétique pour créer des formules qui mutent et

s'améliorent au fur et à mesure, mais c'est une recherche trop colossale pour un stage de deux mois...

Plus modestement, nous allons essayer toutes les lois d'une complexité donnée, et les comparer entre elles.

5.2 Arbres de loi

L'ensemble des arbres de loi que nous considèrerons est le plus petit ensemble qui vérifie :

- k est un arbre de loi ;
- si A est un arbre de loi, l'arbre (r, A), dont la racine r est exp, log, abs ou inv, est un arbre de loi ;
- si A_1 et A_2 sont des arbres de loi, l'arbre (r, A_1, A_2), dont la racine r est $+$ ou $*$, est un arbre de loi.

On pourra s'étonner que les constantes, qui serviront de paramètres à la loi, ne soient pas présentes. L'idée est que chaque expression est définie à une constante additive et une constante multiplicative près. Voyons cela plus en détails.

5.3 Sémantique des arbres de loi

On s'intéresse à la traduction des arbres de loi en expressions algébriques habituelles. Ceci expliquera au passage la sémantique de ces arbres.

Notons $t(A)$ la traduction d'une branche A. On prend :

- k est traduit par k ;
- $(\exp, A)$ est traduit par $\exp(cte * t(A))$;
- $(\log, A)$ est traduit par $\log|t(A) + cte|$;
- (abs, A) est traduit par $|t(A) + cte|$;
- (inv, A) est traduit par $1/(t(A) + cte)$;
- $(+, A_1, A_2)$ est traduit par $t(A_1) + cte * t(A_2)$;
- $(*, A_1, A_2)$ est traduit par $(t(A_1) + cte) * (t(A_2) + cte)$.

Lors de la traduction, les constantes sont numérotées $a_1, a_2, \ldots$ au fur et à mesure qu'elles sont rencontrées. La traduction d'un arbre A est $t(A) + cte$. Nul besoin de constante multiplicative à ce stade, puisqu'elle est imposée par la normalisation.

Notons, par exemple, qu'on n'introduit pas de constante pré-additive dans la traduction de l'exponentielle : en effet, une telle constante serait amalgamée avec la constante multiplicative de l'expression finale.

Avec cette traduction, on peut obtenir, à partir de tous les arbres de lois, toutes les formes de loi obtenues à partir de divers paramètres, de l'exponentielle, du logarithme, de la valeur absolue, de l'inverse et de la multiplication... donc aussi de la division, de la soustraction et des fonctions puissance.

En explorant l'ensemble des arbres de loi, on explore donc un ensemble des lois raisonnables pour F.

5.4 Relation d'ordre sur les arbres de loi

On définit la profondeur $p(A)$ d'un arbre A par :
- la profondeur de k est 0 ;
- la profondeur de (r, A_1) est $1 + p(A_1)$;
- la profondeur de (r, A_1, A_2) est $1 + \max(p(A_1), p(A_2))$.

On définit une relation d'ordre sur les arbres de loi de la façon suivante :
- si les profondeurs sont différentes, l'ordre est l'ordre des profondeurs ;
- sinon, et si les racines sont différentes, c'est l'ordre des racines qui compte, avec par exemple exp $<$ log $<$ + $<$ $*$ $<$ abs $<$ inv ;
- sinon, et si les premières branches sont différentes, c'est l'ordre sur la première branche ;
- sinon, et s'il y a une seconde branche, c'est cet ordre qui compte.

Dorénavant, quand on rencontrera une branche $(+, A_1, A_2)$, on supposera que $A_1 < A_2$. Ceci ne nuit pas à la généralité des traductions possibles des arbres. De même, pour $(*, A_1, A_2)$, on supposera que $A_1 \leq A_2$.

Muni de cette relation d'ordre, l'ensemble des arbres de loi est isomorphe à $\mathbb{N}$. On peut donc le parcourir : tout arbre sera obtenu en temps fini.

5.5 Utilisation des arbres de loi

On se donne une profondeur maximale pour l'arbre qui donnera F. On parcourt l'ensemble des arbres de loi de profondeur inférieure ou égale. Pour chaque loi, on optimise les paramètres pour maximiser $P(\Sigma)$. À la fin, on retient la loi qui a donné le meilleur résultat.

Si on se limite à des arbres de profondeur 1, on retrouve le modèle monopôle en exponentielle évoqué dans la partie précédente pour toutes les expériences réalisées. À profondeur 2, on trouve des lois différentes suivant les votes.

films	$F(k)$	$=$	$(e^{a_1 k} + a_2)((k + a_3)(k + a_4) + a_5) + a_6$
président	$F(k)$	$=$	$e^{a_1(k+a_2)(k+a_3)} + a_4$
mode de scrutin	$F(k)$	$=$	$\lvert e^{a_1 k} + a_2 \rvert + a_3$
actions utiles	$F(k)$	$=$	$\lvert e^{a_1 k} + a_2 \rvert + a_3$
couleur des chaussettes	$F(k)$	$=$	$(e^{a_1 k} + a_2)((k + a_3)(k + a_4) + a_5) + a_6$
pays développés	$F(k)$	$=$	$(e^{a_1 k} + a_2)(e^{a_3 k} + a_4) + a_5$

Pour trouver une loi qui ait un bon pouvoir descriptif dans la plupart des cas, nous allons nous inspirer de ces résultats et étudier les lois suivantes :

$F(k)$	$=$	$e^{a_1 k} + a_2$	(exp)
$F(k)$	$=$	$e^{a_1 k^2} + a_2$	$(gauss)$
$F(k)$	$=$	$e^{a_1 k^2 + a_2 k} + a_3$	(eg)
$F(k)$	$=$	$e^{a_1 k} + a_2 e^{a_3 k} + a_4$	(sum)
$F(k)$	$=$	$(e^{a_1 k} + a_2)(k^2 + a_3 k + a_4) + a_5$	$(epoly)$

À titre comparatif, nous avons aussi utilisé, parmi les lois monopôles de même pôle σ_0, celle qui maximise la probabilité de l'expérience : $\mu(k) = n_k/s$, que nous nommerons (max).

5.6 Étude comparative des différentes lois

Nous avons optimisé chacune des lois précédentes pour chaque expérience de vote : les pays développés (dv), la couleur des chaussettes (ch), les modes de scrutin (sc), les actions utiles (ac), les présidentielles (pr) et les films (fi) . Les résultats sont donnés dans le tableau suivant. Le premier nombre, en petits caractères, est le logarithme (en base 10) de $\mu(\Sigma)$; le second est la valeur de x correspondante (pourcentage d'information capturée par le modèle).

	dv	ch	sc	ac	pr	fi
exp	-125,42 66,33%	-121,05 41,68%	-87,12 19,31%	-106,25 27,42%	-650,15 21,53%	-218,85 4,78%
gauss	-128,73 66,14%	-122,12 40,21%	-86,45 20,61%	-105,97 28,18%	-657,22 20,51%	-218,85 4,78%
eg	-125,41 66,34%	-120,88 41,91%	-86,43 20,65%	-105,93 28,28%	-646,98 21,99%	-218,81 4,80%
sum	-125,35 66,38%	-120,96 41,80%	-87,12 19,31%	-106,25 27,42%	-650,13 21,53%	-218,85 4,78%
epoly	-125,41 66,34%	-120,57 42,34%	-85,64 22,16%	-105,92 28,31%	-647,87 21,86%	-218,76 4,84%
max	-123,53 67,59%	-117,72 46,25%	-82,88 27,47%	-104,86 31,18%	-630,13 24,42%	-213,90 8,24%

Pour l'analyse, rappelons-nous que les lois *exp* et *gauss* sont simples, alors que les lois *eg*, *sum* et *epoly* sont plus complexes : en particulier, elles sont toujours au moins aussi bonnes que *exp* (par un choix de paramètres trivial). Elles ne sont donc intéressantes que si elles améliorent sensiblement le pouvoir descriptif du modèle.

Pour le vote sur les pays développés, c'est *sum* qui est la plus efficace; cependant, elle est à peine meilleure que *exp*, qui est bien meilleure que *gauss* (avec une différence de 3 sur l'entropie, donc un facteur 1000 sur la probabilité).

Pour le vote sur les chaussettes, c'est *epoly*. Cependant, là encore, elle n'améliore que peu la loi *exp*, qui est encore une fois meilleure que *gauss*.

Pour le vote sur les modes de scrutin, c'est encore *epoly*, mais toujours de manière peu convaincante : on gagne seulement 1 sur l'entropie par rapport à la loi gaussienne, qui est beaucoup plus simple. La nouveauté, c'est que celle-ci est mieux placée que *exp*, mais la différence n'est pas énorme (1 sur l'entropie également).

Pour le vote sur les actions utiles, c'est encore une fois *epoly*, mais ce n'est pas si étonnant : cette loi possède beaucoup de degrés de liberté pour l'ajustement au profil expérimental. Mais là encore, le gain par rapport aux profils simples n'est pas très grand. La loi gaussienne est, ici aussi, meilleure que la loi exponentielle, mais de peu.

Pour le vote sur les présidents, c'est le profil mixte *eg*, mais qui n'est pas beaucoup meilleur que *exp* : on peut s'en tenir à une loi simple, de préférence *exp*, qui donne une différence d'entropie de 7 par rapport à *gauss*.

Ainsi, les profils complexes n'apportent que peu de pouvoir descriptif en plus des modèles simples. Parmi ceux-ci, *gauss* est parfois un peu meilleur que *exp*, mais il arrive que *exp* soit bien meilleur que *gauss* : ceci nous incite à garder la loi

exponentielle comme une bonne description générale des votes expérimentaux.

Il n'aura pas échappé au lecteur attentif que la loi exponentielle que nous avons utilisée dans ce chapitre n'est pas la même que précédemment. La constante additive peut être interprétée comme représentant les électeurs sans parti, qui sont uniformément répartis sur $\mathcal{S}_n$: en termes physiques, c'est du bruit.

Chapitre 6

Modèle dipôle

Nous aimerions, à terme, modéliser le vote par une somme de pôles et un bruit. Ce modèle de culture, utilisé dans la littérature, est parfois nommé « société réelle ». Un électeur donné a une certaine probabilité d'être dans chacun des « partis », qui ont chacun un pôle (doxa) et un paramètre p différent. S'il n'est dans aucun parti, il vote de manière équiprobable.

Ce genre de profil semble être confirmé par l'expérience. En effet, observons le graphe du vote sur les présidents, en se limitant aux options {1 JOSPIN, 2 MAMÈRE, 3 BAYROU, 4 BESANCENOT }.

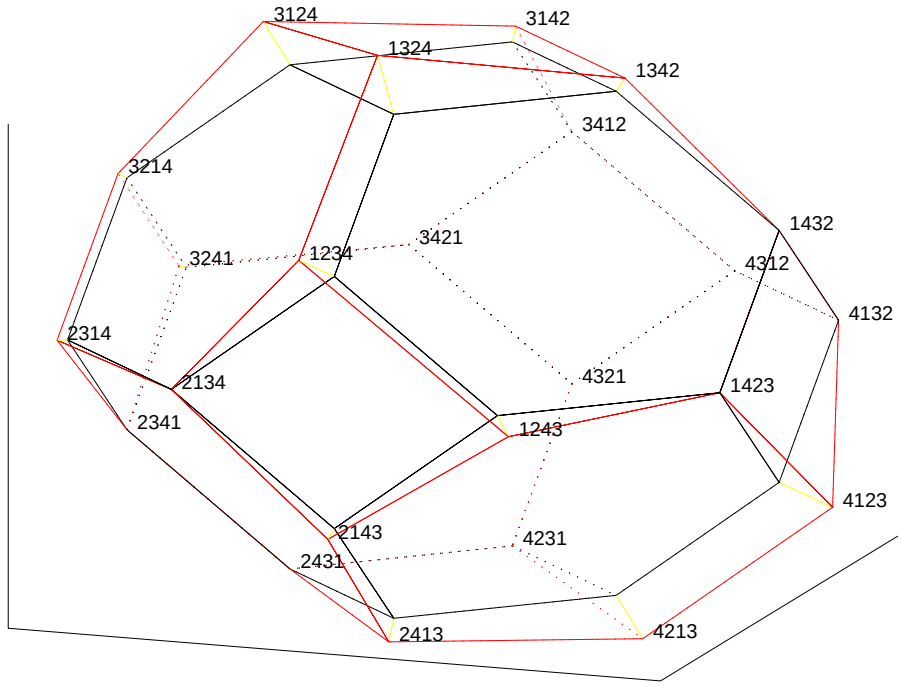

FIG. 6.1: Vote sur les présidents, où 1 = JOSPIN, 2 = MAMÈRE, 3 = BAYROU, 4 = BESANCENOT

Il semblerait qu'il y ait deux pôles, l'un situé sur la face hexagonale où Bayrou (3) est en queue du classement, et l'autre situé en [1 Jospin, 3 Bayrou, 2 Mamère, 4 Besancenot].

Le fait de passer d'un modèle monopôle à un modèle multipolaire pose d'importantes difficultés pour l'optimisation. Pour limiter les problèmes de complexité des algorithmes, nous nous en sommes tenus à deux pôles, mais nous verrons que les résultats sont déjà instructifs.

6.1 Principe du modèle

On suppose que les électeurs votent de manière indépendante, avec la même loi de probabilité :

$$\mu(\sigma) = Ap_1^{d(\sigma,\sigma_0)}/Z_1 + (1-A)p_2^{d(\sigma,\sigma_0')}/Z_2$$

Nous pourrons aussi ajouter une constante additive. Dans ce cas, nous prendrons :

$$\mu(\sigma) = (1-B)(Ap_1^{d(\sigma,\sigma_0)}/Z_1 + (1-A)p_2^{d(\sigma,\sigma_0')}/Z_2) + B/n!$$

Comment optimiser σ_0 et σ'_0 ? Contrairement au cas monopôle, il n'y a *a priori* pas de condition indépendante de p_1, p_2 et A. Et bien sûr, parcourir $\mathcal{S}_n^2$, en optimisant les paramètres à chaque fois, pour trouver les deux pôles, est impensable en terme de complexité.

6.2 Algorithme

Nous avons donc, *a priori*, renoncé à un algorithme 100% sûr pour déterminer les deux pôles.

Condition de départ : on se donne deux permutations.

Boucle conditionnelle : on laisse un pôle fixe (disons f). On optimise les paramètres de la loi dipôle pour toutes les permutations à distance 1 de l'autre pôle (disons m), puis on choisit celle qui donne la meilleure probabilité (m'), si celle-ci est meilleure que la proba initiale : on suit donc la ligne de plus grande pente.

Si m' est meilleur que m *alors* on recommence avec m = m' ;

Sinon, on laisse fixe le pôle m et on s'attaque à f.

Condition d'arrêt : si deux échecs consécutifs interviennent, c'est qu'on est bloqué en bas d'une vallée, tant pour le pôle 1 que pour le pôle 2 : l'algorithme est alors terminé.

La question qui reste à résoudre est : comment choisir les deux pôles de départ. Il est loin d'être sûr — et c'est infirmé expérimentalement — qu'un maximum local soit un maximum global.

La première idée est de prendre chaque couple (σ_i, σ_j), où $\Sigma = (\sigma_1, \ldots, \sigma_s)$ est la suite des votes, d'appliquer à chaque fois l'algorithme, et de garder le meilleur résultat. En effet, on a du mal à croire qu'un maximum de la probabilité se trouve dans une zone où la fonction nombre de votes est nulle. On espère donc que, par cette méthode, on va trouver le maximum global. Ce n'est qu'un espoir, mais il serait intéressant de prouver ce fait ou d'arriver à trouver un contre-exemple.

La seconde idée est de prendre le pôle optimisé du modèle monopôle et chaque σ_i, et de procéder de la même façon.

À chaque fois que nous avons comparé les deux méthodes, elles donnaient les mêmes pôles : ceci est très intéressant car la seconde, *a priori* moins bonne (car explorant moins de points de $\mathcal{S}_n^2$), est beaucoup plus rentable du point de vue complexité. Théoriquement, le facteur entre les deux méthodes est de l'ordre de $(s+1)/2$, où s est le nombre de votants ; en pratique, on a constaté des facteurs allant de 10 à 25.

6.3 Résultats

6.3.1 Vote sur les pays développés

Modèle sans bruit

pôle 1	1 2 3 4 5 6 8 7
pôle 2	1 2 5 4 3 6 7 8
p_1	$0,22$
p_2	$0,33$
A	$0,60$
x $(\log(\mu(\Sigma)))$	$73,32\%$ (-114,85)

Modèle avec bruit

pôle 1	1 2 3 4 5 6 8 7
pôle 2	1 2 5 4 3 6 7 8
p_1	$0,22$
p_2	$0,29$
A	$0,62$
B	0.02
x $(\log(\mu(\Sigma)))$	$73,58\%$ (-114,46)

Dans les deux modèles, les pôles sont les mêmes, avec des paramètres p semblables. Il faut dire que, dans le second modèle, le bruit est faible. Par conséquent les valeurs de x sont proches : le modèle avec bruit n'est pas beaucoup meilleur.

Pour mémoire, le pôle du modèle monopôle était 1 2 3 4 5 6 7 8 : ce n'est pas le pôle principal du modèle dipôle mais une « synthèse » des deux.

Ceci dit, on constate une certaine similitude entre les deux pôles. On peut même se demander si le modèle dipôle est adapté, ou s'il faudrait s'orienter vers un modèle où les électeurs ont une certaine « religion » en matière de 3 4 5 et, de façon presque indépendante, une religion en matière de 7 8 : on pense soit à un modèle où les pôles ne concerneraient qu'une partie des options, et où chaque électeur choisirait plusieurs pôles pour se construire une opinion (« en kit » !), soit à un modèle où les pôles ne seraient pas des permutations mais des relations quelconques, éventuellement non transitives.

6.3.2 Vote sur la couleur des chaussettes

Modèle sans bruit

pôle 1	2 1 3 5 4 6
pôle 2	1 2 3 4 6 5
p_1	$0,54$
p_2	$0,35$
A	$0,51$
x $(\log(\mu(\Sigma)))$	$45,18\%$ (-118,49)

Modèle avec bruit

pôle 1	1 2 3 4 6 5
pôle 2	1 2 3 5 6 4
p_1	$0,37$
p_2	$0,25$
A	$0,69$
B	$0,18$
x $(\log(\mu(\Sigma)))$	$47,25\%$ (-116,98)

Dans ce cas, le bruit n'est pas négligeable, et d'ailleurs, un seul pôle est commun aux deux modèles : dans le modèle sans bruit, le premier pôle prend en charge une tendance locale mais aussi une partie du bruit : c'est pourquoi il est plus éloigné de l'autre pôle que dans le modèle avec bruit.

Le modèle avec bruit apporte, pour ce vote, une amélioration significative à la description, d'environ 2%.

Là encore, le pôle du modèle monopôle, 1 2 3 4 5 6, est un compromis entre les deux pôles du modèle dipôle.

6.3.3 Vote sur les actions utiles

Modèle sans bruit

pôle 1 pôle 2	3 1 2 4 5 2 5 1 4 3
p_1 p_2 A	0,38 0,60 0,58
x $(\log(\mu(\Sigma)))$	38,47% (-102,15)

Modèle avec bruit

pôle 1 pôle 2	3 1 2 4 5 2 1 4 5 3
p_1 p_2 A B	0,32 0,44 0,64 0.26
x $(\log(\mu(\Sigma)))$	38,90% (-101,99)

Curieusement, ici, le bruit est non négligeable mais le modèle avec bruit n'est pas bien meilleur que l'autre. En fait, comme dans le vote sur les chaussettes, l'éloignement du pôle secondaire par rapport au pôle primaire dans le modèle sans bruit lui permet d'émuler une partie du bruit.

Le pôle unique était 1 2 3 4 5 : encore une synthèse, et non pas un des deux pôles. Par exemple, « élever un enfant », qui arrive respectivement en premier et dernier dans les deux pôles, est en milieu de classement dans le pôle unique.

6.3.4 Vote sur les modes de scrutin

Modèle sans bruit

pôle 1 pôle 2	1 2 3 4 5 6 4 3 1 2 5 6
p_1 p_2 A	0,66 0,08 0,82
x $(\log(\mu(\Sigma)))$	27,36% (-82,94)

Modèle avec bruit

pôle 1 pôle 2	1 2 3 4 5 6 4 3 1 2 5 6
p_1 p_2 A B	0,45 0,09 0,75 0,27
x $(\log(\mu(\Sigma)))$	27,36% (-80,38)

Malgré un bruit important, le modèle avec bruit est à peine meilleur que le modèle sans bruit. Dans ce dernier, la part de bruit n'est, dans ce cas, pas émulée par un déplacement du pôle secondaire mais par un élargissement de la loi exponentielle correspondant au pôle principal.

Celui-ci est d'ailleurs le même que dans le modèle monopôle, mais ceci n'est pas étonnant vue la faible importance du second pôle (paramètre A élevé).

6.3.5 Vote sur les présidents

Modèle sans bruit

pôle 1	1 3 2 7 4 6 5 10 8 12 9 13 14 11 15 16
pôle 2	1 9 5 6 8 2 3 11 4 13 14 7 10 12 15 16
p_1	$0,56$
p_2	$0,71$
A	$0,54$
x $_{(\log(\mu(\Sigma)))}$	$26,32\%$ $_{(-616,95)}$

Modèle avec bruit

pôle 1	1 3 7 4 2 10 6 5 8 12 9 13 14 11 15 16
pôle 2	1 5 2 6 9 3 8 11 4 13 14 7 10 12 16 15
p_1	$0,55$
p_2	$0,65$
A	$0,50$
B	$0,04$
x $_{(\log(\mu(\Sigma)))}$	$27,85\%$ $_{(-606,34)}$

Ici, il y a peu de bruit mais l'information gagnée est non négligeable, surtout en valeur absolue car cette expérience contient beaucoup d'information. Les pôles trouvés ne sont pas les mêmes, mais on constate des similitudes.

Sur les limites du modèle dipôle, les remarques que nous avons faites sur les pays développés s'appliquent ici. Dans le modèle avec bruit, on constate par exemple que, si les deux partis sont d'accord pour mettre 15 et 16 en queue, ils diffèrent par leurs positions mutuelles. Mais est-ce que ceci est fortement lié, en pratique, aux positions mutuelles de 5 et 6 ? On pense de nouveau à des modèles alternatifs.

6.3.6 Vote sur les films

Modèle sans bruit

pôle 1	1 2 6 3 5 7 4 8
pôle 2	4 8 5 3 2 1 7 6
p_1	$0,71$
p_2	$0,59$
A	$0,74$
x $_{(\log(\mu(\Sigma)))}$	$9,55\%$ $_{(-212,03)}$

Modèle avec bruit

pôle 1	1 2 6 3 5 7 4 8
pôle 2	4 8 5 3 2 1 7 6
p_1	$0,71$
p_2	$0,59$
A	$0,74$
B	$0,00$
x $_{(\log(\mu(\Sigma)))}$	$9,55\%$ $_{(-212,03)}$

Dans ce vote, le bruit trouvé est quasi-nul : c'est étonnant car, *a priori*, ce vote est celui qui résiste le plus aux analyses non symétriques. Le modèle dipôle, tout comme le modèle monopôle, ne rencontre ici qu'un succès mitigé, mais ce n'est pas vraiment un vote d'opinion, et la culture, sans être désordonnée (l'échec évident d'un modèle uniforme le prouve), est sans doute plus complexe.

Le modèle monopôle donnait 1 2 3 4 5 6 7 8 : encore une fois, cet ordre de préférence est en quelque sorte une moyenne des deux pôles trouvés ici.

6.4 Conclusions communes

Nous aurions pu envisager d'ajouter un pôle en gardant le pôle déjà trouvé (par le modèle monopôle), et en ajoutant un second pôle, considéré *a priori* comme petit devant l'autre, qui optimise la vraisemblance. En fait, c'est une mauvaise idée : l'estimation directe d'un modèle dipôle prouve qu'il n'y a pas un pôle principal prépondérant et un pôle secondaire d'une importance nettement inférieure, d'au moins un ordre de grandeur : le paramètre A n'est proche ni de 0 ni de 1 (d'une distance petite devant 1).

En revanche, le modèle dipôle que nous avons optimisé ici apporte en général une amélioration conséquente du pouvoir descriptif du modèle, par rapport au monopôle : l'annexe A, qui présente une synthèse des résultats, le prouve.

On pourra s'étonner de la faiblesse des coefficients x trouvés, représentant la part d'information captée par le modèle : environ 30% en général. Pour la comparaison, nous avons simulé une expérience fictive, à 100 votants et 6 options. Les deux pôles étaient [1 2 3 4 5 6] et [3 2 5 4 1 6] avec des paramètres $0,6$ et $0,5$. Le paramètre A, proportion du premier parti, valait $0,60$.

L'algorithme a effectivement trouvé les bons pôles, avec les paramètres $0,55$ et $0,52$, et $A = 0,66$. Mais surtout, l'indicateur x valait $32,44\%$, ce qui est proche de ce que nous trouvons pour les profils expérimentaux.

L'information nécessaire pour décrire une réalisation du vote (suivant la loi théorique) correspond en partie à la loi sous-jacente, et en partie à l'écart entre la réalisation elle-même et la moyenne théorique. Il est donc normal que le modèle ne capte que la première partie : les pourcents d'information manquants (ici, 67,56%) correspondent donc à l'écart de la réalisation par rapport à la loi de probabilité sous-jacente.

Ceci montre que nos valeurs de x sont en fait très bonnes, et que le modèle développé décrit bien la réalité.

Cependant, les similitudes et les différences entre les deux pôles trouvés pour chaque expérience semblent suggérer que le modèle dipôle n'est pas le plus naturel pour décrire la structure de l'opinion : on pense par exemple à des modèles de pôles « en kit » ou des modèles où le pôle n'est pas une permutation mais une relation binaire quelconque.

Conclusion

Nous avons réalisé un site web de vote online, qui a assez bien fonctionné puisque nous avons eu un nombre de votes raisonnable pour une analyse statistique. Nous avons fait correspondre aux profils expérimentaux un modèle monopôle, en loi exponentielle par rapport à la distance de la différence symétrique, qui capte en général une partie non négligeable de l'information nécessaire pour décrire le vote. Ensuite, nous nous sommes interrogés sur la pertinence de cette loi exponentielle pour le modèle monopôle. Grâce aux arbres de loi, nous avons montré que la loi exponentielle réalise un bon compromis entre pertinence du modèle et complexité de l'expression.

Une fois l'hypothèse exponentielle confortée, nous avons développé un modèle dipolaire et résolu de nouveaux problèmes algorithmiques. Nous avons renoncé à trouver un algorithme infaillible, mais celui que nous utilisons semble donner des résultats très satisfaisants. Le modèle dipôle présente un pouvoir descriptif significativement supérieur, surtout quand on lui adjoint une composante constante censée représenter les électeurs « sans parti » ou, en termes physiques, le bruit.

Cependant, les rapports entre les deux pôles trouvés pour chaque expérience nous amènent, pour le futur, à nous orienter vers des modèles différents du dipôle pour décrire les votes, en utilisant des pôles composites ou des pôles qui seraient des relations binaires quelconques, voire probabilistes. On peut aussi penser à de nombreuses autres méthodes d'analyses, comme l'étude des corrélations entre les comparaisons par paires ou la séparation en composantes principales.

Les modèles que nous avons utilisés donnent malgré tout une description relativement fidèle des expériences effectuées, et permettent, pour des travaux ultérieurs, de générer des profils de préférences plausibles. On pense en particulier à les utiliser pour évaluer la probabilité, pour un mode de scrutin donné, de se trouver dans des situations paradoxales. On espère ainsi pouvoir comparer le comportement de divers modes de scrutin dans des conditions réalistes.

Remerciements

Je remercie tout d'abord Jean-François pour avoir répondu par mail à mes questions, environ un an avant cette recherche, quand j'ai commencé à me pencher sur la théorie du choix collectif, puis, au cours de ce stage, pour sa disponibilité, malgré nos contraintes respectives, sa compétence et la pertinence de ses conseils.

Je remercie tous les amis qui m'ont aidé. Merci à B̆akkhûs, Pierre-Andrew, Chewie, Roupoil, Aïvonne et Klöcky pour nos discussions fructueuses. Merci à Jérôme Plût et Sarah Loyer pour m'avoir aidé à faire connaître mon site. Merci à David Monniaux pour ses conseils et son soutien.

J'adresse un gros merci à Axelle pour son support logistique, et j'exprime ma reconnaissance éternelle à Julien Reynier et Fabien Mathieu pour leur aide spirituelle.

Un énorme merci à Mathilde pour avoir supporté ces vacances studieuses malgré un départ imminent, écouté mes bêtises et même fait l'effort de les comprendre ! Merci pour son aide, tout au long de l'été.

Enfin, merci à tous les internautes qui ont pris le temps de participer à l'expérience et de permettre ainsi cette étude.

Annexe A

Tableau synthétique des résultats

Nous avons reporté ici les résultats des différents modèles pour chaque expérience de vote : les pays développés (dv), la couleur des chaussettes (ch), les modes de scrutin (sc), les actions utiles (ac), les présidentielles (pr) et les films (fi). Le premier nombre indiqué, en petits caractères, est le logarithme (en base 10) de $\mu(\Sigma)$; le second est la valeur de x correspondante (pourcentage d'information capturée par le modèle).

Les modèles monopôles et dipôles sont ceux en loi exponentielle. Le modèle « monopôle optimal » est le modèle qui utilise le même pôle que celui en loi exponentielle, mais en ajustant $\mu(k)$ au mieux de façon numérique. Le modèle idéal est celui où la probabilité d'une permutation est sa fréquence observée expérimentalement : c'est le modèle qui optimise $\log(\mu(\Sigma))$.

	dv	ch	ac	sc	pr	fi
uniforme	-225,67 0%	-151,44 0%	-116,43 0%	-97,15 0%	-799,24 0%	-225,67 0%
monopôle sans bruit	-125,85 66,05%	-121,68 40,81%	-107,80 23,26%	-88,89 15,90%	-667,03 19,09%	-218,88 4,75%
monopôle avec bruit	-125,42 66,33%	-121,05 41,68%	-106,25 27,42%	-87,12 19,31%	-650,15 21,53%	-218,85 4,78%
monopôle optimal	-123,53 67,59%	-117,72 46,25%	-104,86 31,18%	-82,88 27,47%	-630,13 24,42%	-213,90 8,24%
dipôle sans bruit	-114,85 73,32%	-118,49 45,18%	-102,15 38,47%	-82,94 27,36%	-616,95 26,32%	-212,03 9,55%
dipôle avec bruit	-114,46 73,58%	-116,98 47,25%	-101,99 38,90%	-80,38 32,30%	-606,34 27,85%	-212,03 9,55%
idéal	-74,54 100%	-78,52 100%	-79,30 100%	-45,22 100%	-106,69 100%	-82,82 100%

Annexe B

Spécifications de quelques fonctions d'analyse utilisées

Nous avons reporté ici les premières lignes de commentaire des principales fonctions MatLab qui ont servi à l'analyse.

B.1 Détermination de constantes utiles

```
T_N_K   Nombre de permutations ayant k inversions

 t = T_n_k (n_max, k_max)

    n_max   entier strictement positif
    k_max   entier positif

    t est une matrice de taille (n_max, k_max+1).
    t(n, k+1) est le nombre de permutations de n
    éléments ayant k inversions (i-e k couples
    i<j tels que sigma(i)>sigma(j) ).

    Si on veut tous les coefficients pour n <= n_max,
    il faut prendre k_max >= n_max * (n_max - 1) / 2
    (on peut prendre k_max = Inf).

    Alors, la somme de t sur la n-ième ligne vaut
    factorielle(n).

    N'UTILISE aucune sous-fonction.

    VOIR rien du tout.

CONST_NORM   Constante de normalisation du modèle monopôle en
exponentielle
```

```
 Z = const_norm (p, n_opt, t)

    p   paramètre du modèle monopôle (0 <= p <= 1)
    n_opt   nombre d'options du vote
    t   matrice des T_n_k

    Z est la constante de normalisation du modèle monopôle où

       mu(sigma) = (1/Z) * p ^ d(sigma, pole).

    N'UTILISE aucune sous-fonction.

    VOIR T_n_k.
```

B.2 Génération aléatoire de votes monopôles

```
RAND_DATA_MONOPOLE Matrice de données aléatoire suivant une loi
monopôle

 data = rand_data_monopole(n_vot, pole, p)

    n_vot   nombre de votants
    pole   permutation pôle du modèle
    p   paramètre du modèle

    data est la matrice de données d'une expérience de
    n_votes votes, simulée grâce à rand_perm_monopole, qui génère
    des votes aléatoires en utilisant un modèle monopôle.

    Elle est au même format que les matrices de données :
    data(v, I) est le rang donné à l'option I par l'électeur v.

    Pour tester la validité des résultats obtenus (et ainsi
    vérifier que la fonction permut_monopole est crédible),
    on suggère d'utiliser n_occur_ligne.

    UTILISE rand_perm_monopole, sigma_convert.

    VOIR permut_monopole, n_occur_ligne.
```

B.3 Optimisation du pôle unique

```
ANTI_PERMUT_OPTIMALE Permutation qui optimise la somme
sur-diagonale

 sigma = anti_permut_optimale (M)

    M   matrice antisymétrique
```

```
   sigma est une permutation des indices telle que la somme des
   coefficients sur-diagonaux de M soit maximale.

   UTILISE anti_grp_Cond, anti_perm_opt_glouton.

   VOIR rien du tout.
```

B.4 Modèle monopôle en loi exponentielle

```
MONOPOLE_OPTIMISE Optimisation des paramètres d'un modèle monopôle

 [pole, p, log_proba] = monopole_optimise (data)

   data   une matrice de données : data(v,I) est
   le rang donné à l'option I par l'électeur v.

   pole est la permutation pôle du modèle : pole(I) est
   l'option placée en Ième position.

   p est le paramètre de modèle optimisé.

   log_proba est le log10 de la probabilité de l'expérience dans
   le modèle optimisé.

   UTILISE monopole_optimise_pole, monopole_optimise_p.

   VOIR monopole_log_proba, monopole_trace_log_proba.
```

B.5 Recherche de lois pour le modèle monopôle

```
EXPTREE_OPTIMISE_TOUT Optimisation de la mort qui tue des ours

 [loi, parametres, log_proba] = exptree_optimise_tout (data, pole,
 profondeur_max, t)

   data   matrice de données : data(v,I) est
   le rang donné à l'option I par l'électeur v.

   pole   permutation pôle du modèle : pole(I) est l'option
   placée en Ième position.

   profondeur_max   entier positif

   t   matrice des T_n_k

   loi est la loi de probabilité de pôle pole qui attribue le
   maximum de probabilité à l'expérience décrite par data. Cette
```

loi est choisie parmi celles définies par les arbres de loi de profondeur au plus profondeur_max.

parametres est le vecteur des paramètres de la loi qui optimisent la sauce.

log_proba est le logarithme en base 10 de la probabilité de l'expérience, en supposant cette loi avec ces valeurs des paramètres.

UTILISE data_repart_distances, exptree_profondeur, exptree2expr, expr_optimise_repart, exptree_suivant.

VOIR tout le reste, et en particulier T_n_k.

B.6 Modèle dipôle en loi exponentielle

DIPOLE_OPTIMISE_GLOBAL Optimisation d'un modèle dipôle en partant du modèle monopôle

[pole1, pole2, param, log_proba] = dipole_optimise_global (data, n_param, info, t)

data matrice de données : data(v,I) est le rang donné à l'option I par l'électeur v.

n_param nombre de paramètres du modèle (3 ou 4).

info chaîne de caractères utilisée pour les messages d'information.

t matrice des T_n_k.

dipole_optimise_global prend comme point de départ le pôle du modèle monopôle, et chaque vote de la matrice de données. À partir de ce dipôle, elle descend la ligne de plus grande pente pour essayer d'optimiser le modèle, en suivant la technique de dipole_optimise_local.

On laisse un pôle fixe (disons f). On optimise les paramètres de la loi dipôle pour toutes les permutations à distance 1 de l'autre pôle (disons m), puis on choisit celle qui donne la meilleure probabilité (m'), si celle-ci est meilleure que la proba initiale.

En cas de réussite, on recommence avec m = m'; en cas d'échec, on laisse fixe le pôle m et on s'attaque à f.

Si deux échecs consécutifs interviennent, c'est qu'on est

```
   bloqué en bas d'une vallée, tant pour le pole1 que pour le
   pole2 : l'algo local est alors terminé. Si la proba trouvée
   est meilleure que celles obtenues précédemment, on affiche les
   résultats. Si info est non vide, on affiche
   [info, ' : ... % effectués'].

   Puis on recommence en partant du monopôle et du vote suivant.

   pole1, pole2 sont les meilleurs pôles trouvés.
   param est le vecteur des paramètres optimisés.
   log_proba   est la probabilité correspondante.

   UTILISE monopole_optimise_pole, ordonne_et_compte_lignes,
   sigma_convert, dipole_occur, dipole_optimise_param.

   VOIR dipole_log_proba, dipole_optimise_param,
   dipole_optimise_local, dipole_optimise_global_bourrin.
```

B.7 Calcul de la quantité d'information

```
ENTROPIE_MIN Probabilité dans le mdèle sans information

 log_proba = entropie_min (data)

   data   matrice de données : data(v,I) est
   le rang donné à l'option I par l'électeur v.

   log_proba est le logarithme en base 10 de la probabilité
   de l'expérience, dans le modèle équiprobable.

   Ce modèle est le modèle symétrique sans information.
   Il doit donner une minoration des probabilités
   trouvées avec des modèles non grotesques.

   N'UTILISE aucune sous-fonction.

   VOIR entropie_max, entropie_monopole_max.

ENTROPIE_MAX Probabilité dans le modèle idéal

 log_proba = entropie_max(data)

   data   matrice de données : data(v,I) est
   le rang donné à l'option I par l'électeur v.

   log_proba est le logarithme en base 10 de la probabilité
   de l'expérience, dans le modèle où :

      mu(sigma) = n_sigma / n_vot,
```

```
    n_sigma étant le nombre de votes sigma et n_vot le nombre
    total de votes.

    Ce modèle est celui qui correspond au maximum de vraisemblance ;
    il contient toute l'information sur l'expérience (à part l'ordre
    des votes, bien sûr). Il donne donc une majoration des
    probabilités trouvées.

    UTILISE ordonne_et_compte_lignes.

    VOIR entropie_min, entropie_monopole_max.

ENTROPIE_MONOPOLE_MAX Probabilité dans un modèle monopôle idéal

 log_proba = entropie_monopole_max (data, pole, t)

    data   matrice de données : data(v,I) est
    le rang donné à l'option I par l'électeur v.

    pole   permutation pôle : pole(I) est l'option
    placée en Ième position.

    t   matrice des T_n_k.

    log_proba est le logarithme en base 10 de la probabilité
    de l'expérience, dans le modèle où :

       mu(sigma) = n_k / (n_vot * T_n_k),

    et k = d(sigma, pole), n_k étant le nombre de votes à
    distance k de pole et n_vot le nombre total de votes.

    Ce modèle est celui qui correspond au maximum de vraisemblance
    parmi les modèles monopôles de pôle pole. Il donne donc une
    majoration des probabilités trouvées pour ces modèles.

    UTILISE data_repart_distances.

    VOIR entropie_max, entropie_min.
```