Étudier l'activité parlementaire en l'absence de données de votes individuels

Cyril Benoît*

(very preliminary) draft, I wouldn't cite it if I were you ...

Résumé

Lorsqu'ils s'efforcent de rendre compte du comportement des votants au sein d'organes parlementaires, les chercheurs sont fréquemment contraints de rassembler des données lacunaires, la pratique consistant à enregistrer et à publier les résultats de votes au niveau individuel étant loin d'être généralisée d'une institution à l'autre, voire au sein d'une même institution. Si les acteurs présents et la décision finale sont connus, on ignore généralement tout de la répartition des votes et par conséquent, des motivations, des comportements et des attitudes des votants individuels. Le but de cet article est de proposer un éventail de solutions à ce type de difficultés, en les adaptant spécifiquement au cas de l'activité parlementaire. Pour ce faire, on s'appuie sur un modèle probit multivarié développé par Moritz Marbach (2015, 2017). Se plaçant dans un cadre Bayésien, il propose de dériver les distributions des choix de votes de la distribution stochastique exacte des données de décisions. Dans cette première version d'un article en cours d'écriture, on revient sur les solutions apportées dans la littérature au problème de l'absence de données de votes individuels dans les parlements. Les propriétés du modèles sont ensuite introduites. Une section conclusive passe en revue ses limites et ses perspectives d'application.

*CNRS - Sciences Po. Contact:cyril.benoit1@sciencespo.fr

1 Introduction

Pour appréhender les déterminants et les effets de l'activité parlementaire, les politistes ont eu recours à un large éventail de techniques de recherche et d'approches méthodologiques. Cherchant à « faire parler le parlement » pour « comprendre le gouvernement représentatif », De Galembert, Rozenberg et Vigour (2014) ont par exemple proposé d'étudier le contenu des débats en séances, en montrant la contribution de la parole des acteurs à la construction de l'action publique. Dans un autre registre, Squarcioni (2016) s'est intéressée à la dépendances des députés à leur parti, et aux effets de leur « carrière » sur les différentes manières qu'ils avaient d'exercer leur mandat. Depuis quelques années, les réseaux de collaborations législatifs ont également fait l'objet d'une attention significative dans la littérature, notamment pour rendre compte de façon précise du degré de polarisation d'une chambre, et des conséquences des collaborations entre individus sur l'organisation globale des législatures (Briatte 2016).

Parallèlement à ces développements récents, un domaine d'investigation a suscité un fort engouement au cours des vingt dernières années. Il s'agit du comportement de vote des parlementaires, que le but de l'analyse soit (entre autres questionnements) d'évaluer leur propension à agir conformément à leurs promesses électorales (Scharwz, Schädel et Ladner 2010), de déterminer l'influence de *veto players* sur leurs décisions (Tsebelis 1995), ou encore pour mesurer le poids de groupes politiques supranationaux sur les attitudes des députés européens (Hix 2002). Deux raisons principales expliquent cet intérêt. Le vote des parlementaires constitue en premier lieu un objet d'étude relativement sûr pour qui souhaiterait cerner le fonctionnement d'un parlement. Si ses membres disposent de nombreuses autres capacités pour influencer, voire pour prendre des décisions (notamment dans le cadre de commissions ou de comités), il reste dans la grande majorité des situations l'acte de décision final, ainsi que le moyen le plus usuel à la discrétion d'un

parlementaire pour exprimer ses préférences (Aydelotte 1977, Finer 1987). Du fait de sa centralité, il représente par voie de conséquence une métrique commode pour traiter d'un grand nombre de questions de recherche : discipline qu'un parti est en mesure d'imposer ou non aux politiciens qui le composent (Shugart 1998), influence de clientèles diverses sur les mesures adoptées (Moe et Anzia 2016), capacités de résistance d'un organe législatif à un gouvernement cherchant à limiter ses capacités, etc. Une deuxième raison, d'ordre empirique, justifie l'attention portée à cet objet. Alors que les premiers travaux examinant les pratiques de vote dans les parlements devaient se contenter de données relativement parcellaires, composées pour l'essentiel de comptes rendus de séances − lorsqu'ils étaient disponibles − (par ex. Lowell 1901), les données de vote dans les parlements nationaux ont été, de façon croissante sur la période récente, mises à disposition du public et rendues accessible *via* Internet (Hug 2012). Dans une recherche de transparence, de nombreux parlements (notamment dans de « nouvelles » démocraties) se sont engagés à publier les résultats de votes en leur sein, parfois sous la pression des électeurs ou des partis politiques eux-mêmes (Middlebrook 2003, Carey 2009). Cette tendance a largement contribué à faciliter l'accès des chercheurs à ces données, et au renouvellement des études sur l'activité parlementaire se fondant sur le comportement de vote de leurs membres.

En dépit de cet effort de publicisation, une part significative des données de ce type reste inaccessible, notamment pour certaines procédures qui demeurent, d'un parlement à l'autre, inobservables ou placées sous le sceau du secret. Dans ce contexte, les travaux généralisant des constats formulés à la suite de l'examen de certains votes ont donc de fortes chance d'être biaisés, dans la mesure où le comportement des parlementaires pourrait s'avérer très différent pour les cas où les données sont manquantes ou partiellement disponibles (Hug 2012). Si ce problème a été identifié de manière précoce (Van Doren

1990), la plupart des réponses apportées demeurent perfectibles sur le plan méthodologique. Dans de nombreuses situations, la solution proposée consiste en effet, là où seules les décisions de l'institution sont connues, à examiner le comportement des votants de manière agrégée et à déduire des implications de ce résultat. Une telle stratégie charrie à son tour de nombreux biais que nous exposons ci-après.

A l'aune de ce constat, le but de cet article est de proposer une nouvelle réponse à la difficulté posée par l'examen des comportements de votes dans les parlements en l'absence de données au niveau individuel. Nous nous appuyons sur un modèle statistique développé par Marbach (2015, 2017), que nous adaptons plus spécifiquement au cas de l'activité parlementaire. Se plaçant dans un cadre Bayésien, il repose sur la construction d'une fonction de vraisemblance, une méthode utilisant des marches aléatoires sur des chaînes de Markov (de type échantillonneur de Gibbs) étant ensuite mobilisée pour générer une densité *a posteriori* sur des données augmentées. En langage naturel, cette approche revient à obtenir par un grand nombre de simulations les probabilités associées à la décision de chaque votant individuel d'un échantillon, à partir d'un état de connaissance initial plus ou moins précaire sur la situation examinée.

Le reste de l'article est organisé de la manière suivante. Dans la section (2), nous revenons sur les difficultés méthodologiques auxquelles sont confrontées les travaux cherchant à inférer sur les comportements de vote à partir de données agrégées. Dans la section (3), nous présentons et discutons le modèle de Marbach en revenant notamment sur les difficultés techniques posées par son adaptation au contexte de l'activité parlementaire. Une discussion de ses perspectives d'application est proposée en section (4).

2 Une revue des solutions apportées au problème de l'inférence en l'absence de données de vote au niveau individuel

La plupart des institutions parlementaires ne publie pas, ou ne publie que de manière partielle les résultats des votes au niveau individuel (Saalfeld 1995). C'est notamment le cas lors des scrutins par appel nominal (*roll call vote*), où une position exprimée publiquement par un-e représentant-e ne sera généralement ni enregistrée ni publiée par les services du parlement en question (ex. Assemblée Nationale 2017). D'autres procédures (comme certaines nominations personnelles) sont secrètes, rendant délicate toute tentative visant à mesurer précisément l'effet, par exemple, d'un groupe d'intérêt sur la désignation de tel ou tel individu à un poste stratégique. Cette situation n'est pas propre aux parlements nationaux. De nombreuses organisations internationales (Zamora 1980) et banques centrales (Fry et al. 2000) ne publient pas l'intégralité des votes individuels, notamment pour les décisions prises à main levée. La totalité des « comités » composés de « représentants » approuvant ou rejetant des « propositions » (Black 1958) sont donc potentiellement concernés. Du point de vue de la recherche, la difficulté rencontrée par de nombreux universitaires pour accéder à ces données contraste avec la publication et la numérisation croissante d'une partie de la production de telles entités, qui revendiquent comme principe de légitimation une démarche de transparence à l'égard de leurs audiences (De Fine Licht et al. 2011).

Dans la littérature consacrée aux pratiques de votes dans les institutions parlementaires, deux stratégies peuvent être observées en réponse à cet écueil. La première consiste à ignorer cette contrainte en ne soumettant à l'analyse que les cas où des données de vote au niveau individuel sont disponibles. Si elle permet un examen rigoureux sur le plan

méthodologique, une telle stratégie amène généralement l'analyste à privilégier l'examen de certaines procédures au détriment d'autres, le degré de publicisation des données au niveau individuel étant en effet corrélé au mode de scrutin. A titre d'exemple, les votes par scrutin public ordinaire sont assez fréquemment publiés. En revanche, les données de votes individuels ne sont que rarement voire jamais disponibles pour les scrutins publics par appel nominal et ce, dans la plupart des institutions parlementaires du monde [1]. Une telle situation introduit immanquablement un biais dans l'analyse. Hug (2009) a ainsi pu démontrer que la cohésion des partis dans les scrutins publics par appel nominal différait de la plupart des autres types de vote. La portée des travaux qui n'étudient le comportement des acteurs que pour les scrutins publics ordinaires est donc limitée.

Une seconde option a la préférence de nombreux auteurs. Elle consiste à inférer, dans les cas où les données de vote au niveau individuel ne sont pas disponibles, sur le relevé des votes « agrégés ». En d'autres termes, et si l'on retient cette approche, on se fonde sur une liste de propositions « acceptées » ou « rejetées » (variable dépendante), sachant l'orientation d'un parlement, le mode de scrutin, le nombre de partis représentés, ou d'autres variables (indépendantes) jugées pertinentes. Ce faisant, on substitue une analyse en termes de « décision d'acteurs politiques » par une analyse en termes de « décision d'une institution », ce qui pose trois difficultés méthodologiques majeures (Marbach 2015). La première concerne directement la production de l'inférence. Si l'on examine un jeu de variables indépendantes qui font état de variations d'un acteur à l'autre et d'une proposition à l'autre, on ne peut inclure ces variations dans l'analyse du relevé de votes agrégés. Un écart est par conséquent marqué entre la composition de la variable dépendante et celle des variables indépendantes, ce qui rend la qualité de l'infé-

1. Plusieurs exceptions notables peuvent être mentionnées, comme le Congrès des Etats-Unis, le Parlement Autrichien et le Parlement Suisse au niveau national ; et le Parlement européen au niveau supranational.

rence plus incertaine. Une deuxième difficulté a trait au correctif généralement apporté à une telle configuration, qui peut déboucher sur l'amplification de l'effet d'agrégation mis en cause. Il est en effet possible d'agréger les variations entre acteurs identifiées dans les variables indépendantes afin de mieux les faire correspondre à la construction de la variable dépendante (les décisions agrégées). Pour ce faire, on les « neutralise » en s'appuyant sur les informations fournies par les statistiques descriptives (moyenne, statistique d'ordre). Si cette option présente l'intérêt de faciliter l'analyse, elle ajoute simultanément des biais du fait de l'agrégation des données disponibles pour les variables indépendantes, augmentant du même coup le risque d'erreur écologique (Robinson 1950, King 1997, King, Tanner et Rosen 1999, Marbach 2015). Enfin, et même dans les cas où l'on sait avec certitude que les choix des acteurs ne dépendent que des caractéristiques de la proposition sur laquelle ils doivent se prononcer, le relevé des votes agrégés ne permet jamais de déterminer la probabilité que chaque acteur individuel vote en faveur ou contre une proposition. En d'autres termes, on ne peut établir que les variations de la probabilité qu'une proposition soit approuvée ou rejetée sachant les variations des variables indépendantes. Or, et selon les cas, la probabilité qu'un acteur vote en faveur ou contre une proposition pourra être très différente que la probabilité qu'une proposition soit acceptée ou rejetée, ce qui peut amener le chercheur à surestimer ou à sous-estimer l'effet des variables indépendantes (Carrubba et al. 2006).

Face à ces difficultés, une solution alternative consiste à centrer la construction de l'analyse sur la désagrégation des données de votes disponibles, afin d'estimer les résultats au niveau individuel. Si l'on retient cette troisième option, la démarche change de sens. Il ne s'agit plus d'ignorer, ou de composer avec une décision d'institution, mais de partir de cette situation pour retrouver les choix de chacun des votants impliqués afin d'en dégager les déterminants de façon plus précise. Si ce paramètre est inconnu, on peut

néanmoins s'appuyer sur un état de connaissance (*a priori*) relatif au contexte étudié pour mener à bien cette analyse. En recourant à des méthodes d'échantillonnage à partir d'une distribution de probabilités initiale, des extrapolations et des simulations peuvent être effectuées pour améliorer cette information première, et générer une estimation fiable des probabilités (*a posteriori*) de vote associées à chacun des acteurs impliqués. La mise en œuvre d'une telle opération implique un changement d'orientation méthodologique par rapport à celle privilégiée par les adeptes de l'inférence écologique. On va en effet chercher à déterminer ici la probabilité que chaque votant ait approuvé ou rejeté une proposition sachant que ladite proposition a été approuvée/rejetée [2].

3 Estimer les déterminants du vote à partir des probabilités des décisions individuelles

A notre connaissance, Marbach (2015, 2017) a développé le modèle le plus convaincant pour entreprendre une telle analyse. Nous présentons ses propriétés principales dans cette section, avant de revenir sur ses perspectives d'application et ses limites dans la section suivante − le modèle n'ayant pas été spécifiquement construit pour étudier l'activité parlementaire. A l'exception des incises correspondant aux explications des formules mathématiques ou des choix de modélisation retenus, le contenu de la sous-section (3.2) reproduit de manière simplifiée les équations développées dans Marbach (2015, 2017). On se référera à ces deux références pour obtenir les lemmes correspondant ainsi que le détail des calculs intermédiaires, ignorés ici par souci de clarté.

2. Cette façon de poser le problème de l'inférence est caractéristique de l'approche Bayésienne, par opposition au paradigme Fréquentiste dominant dans la littérature. Pour une introduction, voir (Jackman 2009).

3.1 Un modèle multivarié probit reposant sur une approche Bayésienne

Le modèle présenté peut être rattaché à la famille des modèles structurels Bayésiens [3]. Comme leur nom l'indique, ces modèles adoptent une perspective Bayésienne de l'inférence, dont on peut brièvement synthétiser les principes en suivant la description qu'en proposent Kaplan et Depaoli (2012). Prenons une variable aléatoire Y qui prend une fois réalisée une valeur y. A titre d'exemple, Y peut désigner l'éventail des réponses possibles à un questionnaire. Quand une personne choisit l'une de ces réponses, Y prend une valeur y. Y, en tant que variable aléatoire, n'est jamais observée. C'est donc la distribution de probabilités de Y que nous cherchons à comprendre au travers de ses réalisations de valeur y. Dans ce contexte, θ désigne un paramètre que nous considérons comme reflétant la distribution de probabilités d'intérêt. On cherche à déterminer la probabilité d'observer y étant donné un paramètre inconnu θ, soit $\mathcal{P}(y \mid \theta)$. Le but de l'analyse statistique est par conséquent d'obtenir des estimations du ou des paramètres inconnus à partir des données, soit de la vraisemblance des paramètres, que l'on peut écrire $\mathcal{L}(\theta \mid y)$. L'inférence bayésienne se distingue de cette acception générique en raison de sa conception de la nature du paramètre θ. Dans les approches classiques, θ est inconnu, mais fixé, tandis que pour les statisticiens Bayésiens, θ est lui aussi une variable aléatoire, et possède une distribution de probabilités qui reflète notre incertitude quant à sa valeur réelle. Comme θ et y sont aléatoires, on peut modéliser conjointement la probabilité des paramètres et des données comme une fonction de la distribution conditionnelle des données compte tenu des paramètres, et de la distribution *a priori* des paramètres. En d'autres termes, et plutôt que d'adopter un point de vue « objectiviste » sur θ, les statisticiens Bayé-

3. Pour un exemple de modèles Bayésiens appliqué à l'étude des votes dans les parlements, voir Han (2007). Pour une discussion méthodologique sur cette approche, voir Clinton et Jackman (2004).

siens préfèrent considérer que l'information dont nous disposons nécessairement sur le cas étudié doit informer l'analyse, étant entendu que l'on s'attend à observer avec une chance plus ou moins forte telle ou telle réalisation de y. Ce point de vue (*a priori*) conditionne la contribution des observations au renforcement ou à l'affaiblissement de l'hypothèse initiale. Dans l'exemple du questionnaire mentionné plus haut, les partisans de l'approche Bayésienne considèrent en effet que, si l'on considère *a priori* qu'il y a une chance très forte pour que les répondants choisissent une réponse en particulier, la force probatoire des observations allant dans le sens de cette information ne doit pas être la même que pour celles qui en amoindrissent la validité. C'est précisément ce type de hiérarchisation de la qualité des liens de cause à effet mis au jour que permet l'intégration de l'*a priori*, et qu'interdit l'approche statistique standard. Cette logique peut être décrite de façon différente, pour en faciliter la compréhension. Si Θ désigne l'ensemble des valeurs possibles du paramètre d'une loi de probabilité Π (utilisée pour modéliser la distribution *a priori*), on raisonne comme si le paramètre était aléatoire de loi Π. L'objectif de l'analyse va être, en accumulant des observations y, de remplacer la loi marginale de θ (*a priori*) par la loi conditionnelle de θ sachant y, aussi appelée loi *a posteriori* (Gourieroux et Monfort 1996). La loi *a priori* est déterminée par le chercheur, à partir de sa connaissance de l'objet étudié ou d'une revue de la littérature existante. En collectant des données, son objectif ne va pas être de révéler la vraie valeur du ou des paramètres inconnus, mais d'améliorer son information quant à la probabilité qu'une cause soit reliée à un effet, conditionnellement à un état de connaissance initial.

S'il adopte une perspective Bayésienne de l'inférence, le modèle de Marbach se présente plus globalement sous la forme d'un modèle multivarié probit (MVP) dont la structure a été clairement exposée par Aurier et Mejía (2014). Ces modèles visent à faciliter l'identification des déterminants des choix individuels, qui sont généralement spécifiés

comme étant de nature discrète (choix ou non-choix) et comme étant le reflet d'une variable latente non observable modélisée (probabilité du choix). Les modèles multivariées reposent sur le théorème de Besag (1974) qui affirme qu'il est possible de calculer la probabilité jointe des différents éléments d'un ensemble (Aurier et Mejía 2014). Dans le cas qui nous intéresse, cela revient à dire que dans un contexte de choix multiples de votes (oui, non), chaque vote peut être choisi individuellement, et comme chaque choix dépend des autres, la probabilité jointe du choix des votes peut également être calculée. Des méthodes de type MCMC (*Monte Carlo Markov Chain*, voir Geyer (2011)) sont mises en œuvre pour générer une approximation des distributions multivariées.

3.2 Modèle de Marbach (2015, 2017)

3.2.1 Environnement

On considère un environnement de M membres ($i = 1, \ldots, M$) et de J décisions ($j = 1, \ldots, J$). Le vote d'un membre se présente sous la forme d'une variable aléatoire binaire $y_{ij} \in \{0, 1\}$. Deux options sont donc possibles par votant, à savoir accepter une proposition (*oui*) ou rejeter une proposition (*non*). Le choix de vote est déterminé par un vecteur de K covariables, qui sont données au début de l'analyse. Ce vecteur est noté $\mathbf{x}_{ij}$. Lorsque l'on étudie l'activité parlementaire avec pour toute information un relevé de votes au niveau agrégé, on connaît donc le résultat (binaire) du vote $b_j \in \{0, 1\}$ où $b_j = 0$ si la proposition est rejetée. De même, les covariables sont connues. En revanche, on ignore tout de la répartition individuelle des votes. Pour retrouver cette clef de répartition, on va examiner la façon dont les covariables prédisent le vote au niveau individuel sachant le résultat final au niveau agrégé. Ce problème posé sur un mode très classiquement bayésien est résolu en formulant une hypothèse sur la distribution des votes (*a priori*). A l'aide

des covariables, on va demander à un algorithme de chercher différentes combinaisons de votes jusqu'à obtenir la meilleure approximation possible de cette distribution.

Dans un article préliminaire de 2015, Marbach propose une synthèse de l'environnement dans lequel se situe le modèle (Table 1). L'ensemble des comités, commissions ou assemblées sont couvertes par cette représentation schématique (voir également Broniatowski et al. 2009), même si elle s'applique particulièrement bien aux organes parlementaires.

		Observé		Non-observé
Membre	Décision	Covariable	Résultat	Vote
1	1	$\mathbf{x}_{11}$	b_1	y_{11}
2	1	$\mathbf{x}_{21}$		y_{21}
⋮	⋮	⋮		⋮
M	1	$\mathbf{x}_{M1}$		y_{M1}
⋮	⋮	⋮	⋮	⋮
1	J	$\mathbf{x}_{1J}$	b_j	y_{1J}
2	J	$\mathbf{x}_{2J}$		y_{2J}
⋮	⋮	⋮		⋮
M	J	$\mathbf{x}_{MJ}$		y_{MJ}

TABLE 1 – Une synthèse de l'environnement dans lequel se place le modèle, où M désigne les membres d'un comité et J leurs décisions. Le résultat de vote observé (b_j) suit une règle de vote quelconque. Les votes non-observés sont désignés par (y_{ij}), et il existe un vecteur de covariables ($\mathbf{x}_{ij}$) pour chaque combinaison de type membre-décision. Tableau adapté de Marbach (2015)

3.2.2 Enoncé du modèle (Marbach 2015)

Dans ce modèle, l'ensemble des covariables pour la totalité des membres M pour chaque décision j est désigné par une matrice de forme $M \times K$ désignée par $\mathbf{X}_j$. $\mathbf{y}_j$ est le vecteur de norme M regroupant tous les votes, $y_{1j}, \ldots, y_{Mj}$, pour la proposition

correspondante (désigné ci-après « profil de vote »). $\mathbf{y}_j^*$ désigne le vecteur des utilités latentes de M membres correspondant au vote en faveur d'une proposition j. L'unité de ce vecteur est donc l'utilité d'un membre i, notée y_{ij}^*. Un membre i vote *oui* si $y_{ij}^* \geq 0$. Par souci de simplification, on considère ici que l'utilité latente est une fonction linéaire des covariables avec le vecteur du paramètre correspondant β (Marbach 2015).

On part du postulat que la règle de vote qui détermine l'adoption ou le rejet d'un proposition est une *q-rule*. Une règle de vote suit une *q-rule* si une décision est prise si au moins q individus (du total n) votent en sa faveur, soit typiquement : $q > \frac{n+1}{2}$. Le seuil de majorité est noté $\mathcal{R}$. La proposition est rejetée ($b_j = 0$) si le nombre de votes $\sum_{i=1}^{M} y_{ij}$ est inférieur à $\mathcal{R}$. Dans le cas contraire, la proposition est adoptée ($b_j = 1$). Ces éléments de notation permettent à Marbach d'écrire le modèle de la façon suivante :

$$
\begin{aligned}
&\mathbf{y}_j^* = \mathbf{X}_j\beta + \epsilon_j \\
&\epsilon_j \overset{iid}{\sim} \phi(0, \mathbf{1}) \\
&y_{ij} = \begin{cases} 0 & \text{si} \quad y_{ij}^* < 0 \\ 1 & \text{autrement} \end{cases} \\
&b_j = \begin{cases} 0 & \text{si} \quad \sum_{i=1}^{M} y_{ij} < \mathcal{R} \\ 1 & \text{autrement} \end{cases}
\end{aligned}
\tag{1}
$$

où $\phi(0, \mathbf{1})$ est la fonction de densité normale multidimensionnelle. Il convient de noter que ce modèle, à l'instar des modèles probit multivariés présentés dans la section précédentes, repose sur deux postulats implicites. On assume en premier lieu que les coefficients sont partagés par l'ensemble des membres du comité. Le modèle est donc indiqué pour étudier un comité, une commission ou une assemblée où les membres sont

soumis aux mêmes déterminants et sont soumis à la même règle de vote. Cela n'exclue pas, bien entendu, que l'effet de certains déterminants puisse être nul au niveau individuel. Le second postulat est que les choix de votes sont conditionnellement indépendants. Il en va de même pour les propositions soumises à l'appréciation des membres. Ce second postulat, ainsi que le souligne Marbach, est très classique dans les modèles de type *ideal-points*. A nouveau, il est en outre très cohérent avec la notion de discretion que l'on retrouve dans les modèles probits multivariés.

3.2.3 Vraisemblance et densité *a priori* (Marbach 2015, 2017)

Dans ce modèle, la probabilité d'observer une décision équivaut à la somme des probabilités des profils de votes qui ont pu générer la décision en question. La probabilité de chacun de ces profils de votes est le produit des probabilités des choix individuels, qui se présente comme une fonction linéaire de covariables et de paramètres. On retrouve à nouveau ici une propriété essentielle des modèles probit multivariés tels que décrits dans la section précédente. Le produit sur l'ensemble des probabilités de décision permet d'obtenir la vraisemblance des données. Ensuite, on définit la probabilité d'un profil de vote et on établit les profils hypothétiques du vote qui peuvent générer une décision.

En suivant le postulat d'indépendance des choix, la probabilité d'observer un profil de vote $\mathbf{y_j}$ est le produit des probabilités de choix individuels pour une proposition j ou, de façon équivalente, l'intégration de l'utilité latente dans chaque dimension sur l'intervalle correspondant au choix de vote observé. Plus formellement (Marbach 2015, 2017) :

$$\begin{aligned} f(\mathbf{y_j}, \mathbf{X_j}|\beta) &= \int_{p1j} \dots \int_{pMj} \phi(\mathbf{y^*}_j|\mathbf{X}_j\beta) d\mathbf{y^*}_j \\ &= \Phi_{\mathcal{P}(\mathbf{y}_j)}(\mathbf{X}_j\beta), \end{aligned} \tag{2}$$

où $\phi(.)$ est la densité normale multivariée de dimension M et p_{ij} est l'intervalle qui correspond au choix de vote y_{ij} pour le profil $\mathbf{y}_j$, soit $p_{ij} = (0, \infty)$ si $y_{ij} = 1$ et $p_{ij} = (-\infty, 0)$ si $y_{ij} = 0$.

Soit $\tilde{\mathbf{y}}$ un profil de vote hypothétique et soit $V(1)$ un ensemble de profils de votes hypothétiques satisfaisant $\sum_i \tilde{y}_i \geq \mathcal{R}$. Cet ensemble contient donc tous les profils de vote qui réalisent l'adoption d'une proposition ($b_j = 1$). En suivant la même règle, on peut désigner par $V(0)$ un ensemble réalisant le rejet d'une proposition. Chaque ensemble est fini mais peut être potentiellement très large.

A partir de ces définitions, on peut écrire la probabilité pour $b_j = 1$ (le cheminement étant identique pour $b_j = 0$) comme la somme sur les probabilités pour tous les profils de vote hypothétiques qui peuvent réaliser $b_j = 1(b_j = 0)$. La vraisemblance est obtenue en prenant le produit sur l'ensemble des décisions. Formellement :

$$\mathcal{L}(\beta|\mathbf{X}, \mathbf{b}) = \prod_j \sum_{\tilde{\mathbf{y}} \in V(b_j)} \left[\Phi_{\mathcal{P}(\tilde{\mathbf{y}})}(\mathbf{X_j}\beta) \right]. \tag{3}$$

L'approche Bayésienne complète la vraisemblance avec une densité *a priori* pour les paramètres (les coefficients). Dans son modèle, Marbach postule qu'ils suivent une loi normale avec une moyenne *a priori* $\mathbf{b_0}$ et une matrice de covariance diagnonale $\mathbf{B_0}$. La densité *a posteriori* est proportionnelle au produit de la fonction de vraisemblance telle qu'exposée dans l'équation (3) et de la densité *a priori*.

3.2.4 Calcul de la loi *a posteriori*

Comme dans la plupart des modèles bayésiens, l'identification des choix de votes au niveau individuel est ici donnée par l'obtention de la densité de la loi *a posteriori*. Pour ce

faire, on va combiner l'information apportée par l'*a priori* (hypothèse de normalité de la distribution des coefficients) et celle apportée par l'analyse des variables observables. En dépit de son apparente complexité, cette idée peut aisément être décrite en langage naturel. Notre but est d'obtenir une estimation précise de la probabilité que chaque membre d'un comité ait approuvé ou rejeté une proposition. On ne dispose comme information que de la décision finale au niveau agrégée. Dans le cadre de ce modèle, on commence par établir un *a priori* sur la distribution des votes. L'option la moins risquée consiste à affirmer que les votes sont normalement distribués, au sens de la loi normale (voir figure 1), bien qu'il ne s'agisse pas là d'une nécessité.

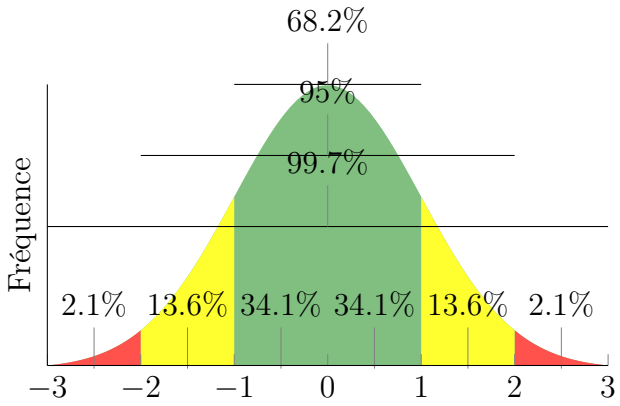

FIGURE 1 – Un exemple de distribution gaussienne, d'après John Canning.

Bien qu'elle ne soit pas encore démontrée, cette affirmation constitue déjà une première idée de la répartition des votes, et donc de la chance relative pour que tel ou tel membre ait voté pour ou contre la proposition examinée. Dans un second temps, on va chercher à affiner ce constat en examinant une série de variables. La loi *a posteriori* du paramètre va être obtenue en combinant l'information initiale avec celle tirée de l'analyse des données. Comme indiqué plus haut, ce résultat va se présenter sous la forme d'une distribution de probabilités sur θ.

Un échantillonneur de Gibbs va être utilisé pour mener à bien cette opération. Il s'agit d'une méthode très populaire qui, en raison de ses propriétés, s'applique particulièrement bien au problème posé par l'identification des probabilités des votes au niveau individuel quand on ne connaît que la décision au niveau agrégé. On trouvera une introduction détaillée à cette approche dans Gordon et Belenger (1996). Comme son nom l'indique, il s'agit d'une méthode d'échantillonnage à partir d'une distribution de probabilités initiale. Des vecteurs vont être générés aléatoirement sur cette distribution de façon itérative, en suivant la séquence déterminée. On choisit des valeurs de départ pour les coefficients ; puis, et conditionnellement à ces valeurs, aux covariables, à la décision connue, des profils de vote sont générés pour toutes les décisions. Conditionnellement aux profils de vote et aux covariables, on génère le vecteur des utilités latentes pour toutes les décisions ; puis, conditionnellement aux utilités latentes et aux covariables, on génère les coefficients. L'opération est répétée jusqu'à obtenir une convergence (voir pour une discussion Marbach 2017).

L'objectif de l'échantilloneur est donc de combiner les informations dont nous disposons pour nous permettre d'inférer sur les données manquantes. Comme l'indiquent les différentes étapes de cette séquence, on part de notre *a priori* sur le type de distribution de l'échantillon (étape 1), que l'on combine à l'information dont on dispose (étape 2) pour trouver le vecteur d'utilité pour chaque décision (étape 3). Ensuite, on cherche à retrouver les coefficients réels à partir des résultats obtenus pour les calculs effectués lors des étapes 2 et 3 (étape 4). En répétant cette opération pour la totalité de l'échantillon, on finit par obtenir une représentation assez précise des probabilités conditionnelles associées à chaque profil de vote [4].

4. Le détail des opérations à effectuer est disponible dans Marbach (2017). On y trouvera également des applications du modèle à partir de données de seconde main, ainsi qu'une discussion sur les coûts d'agrégation et les enjeux d'identification dans ce modèle.

Formellement, on va donc chercher à obtenir une réalisation du paramètre $\theta = (\theta_1, \ldots, \theta_m)$ suivant la loi *a posteriori* à partir des lois conditionnelles. Les itérations de l'algorithme vont progressivement générer les états d'une chaîne de Markov. Après un grand nombre d'itérations, le vecteur θ obtenu est considéré comme la réalisation de la loi *a posteriori* (Robert 2001).

4 Discussion

Les premières tentatives d'application de ce modèle sont prometteuses. En « cachant » volontairement une partie des données mobilisées dans de précédents travaux, Marbach est ainsi parvenu à identifier avec une faible marge d'erreur les probabilités conditionnelles de votes dans le contexte de décisions de la Cour suprême des Etats-Unis et du Conseil de sécurité de l'ONU. Il est également à noter que le coût d'entrée technique lié à sa mise en œuvre est significativement plus réduit que ce que pourrait laisser croire la présentation qui en a été faite dans la section précédente − un R-package très simple d'utilisation (consilium) ayant en effet été développé par son auteur. Il n'est toutefois pas exempt de fragilités, qui peuvent limiter ses perspectives d'application, notamment pour étudier le fonctionnement des institutions parlementaires.

Le modèle charrie tout d'abord les défauts inhérents aux modèles Bayésiens en général, et à ceux recourant à des méthodes MCMC en particulier. En raison de sa contribution importante à la production de l'inférence, l'*a priori*, s'il est mal spécifié, peut aisément gréver l'interprétation des données (Figure 2) − étant entendu que les données agissent comme un médiateur entre la loi *a priori* et la loi *a posteriori*.

Cela se traduit en pratique par le fait que le modèle sera plus performant si l'on dispose d'une meilleure information pour établir la loi *a priori*, tels que les pourcentages

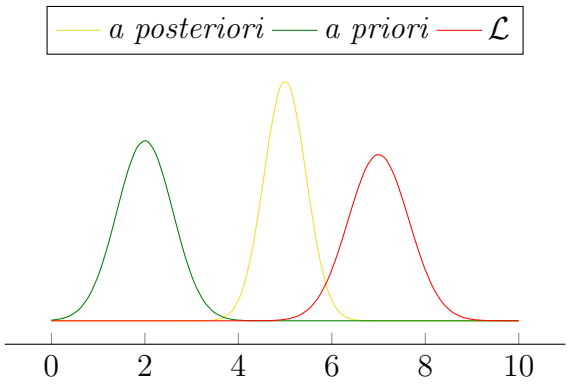

FIGURE 2 – Un exemple d'*a posteriori* peu informatif, en raison d'un *a priori* mal spécifié. Ici, l'*a priori* est donné par $\mathcal{N}(\mu = 1.5, \sigma = 0.4)$ et la vraisemblance par $\mathcal{N}(\mu = 6.1, \sigma = 0.4)$. Le résultat est mathématiquement correct, mais difficilement exploitable en l'état.

des voix exprimées par différents groupes parlementaires, des résultats provenants de recherches antérieures qui examineraient une législature ayant la même composition que celle étudiée, ou encore pour laquelle on disposerait de quelques données de vote au niveau individuel. Cette contrainte est loin d'être insurmontable, mais elle peut, dans un certain nombre de cas très précis, limiter notre capacité à inférer sur les données. Au vue de sa construction, l'utilisation de ce modèle se révèle par conséquent plus sûre pour les cas où l'on dispose d'une information partielle : on pourrait par exemple imaginer une situation où on ne connaîtrait qu'une partie des données de vote au niveau individuel au sein d'une assemblée, et où on chercherait à déterminer les votes d'une autre partie.

Au-delà de cette difficulté, qui peut être contournée de plusieurs manières [5], le modèle est limitée dans son application aux institutions parlementaires du fait de la structure de la méthode MCMC sur lequel il repose. Si l'algorithme utilisé est efficace pour générer

5. Il est par exemple possible de collecter des données qualitatives pour améliorer notre état de connaissance initial sur la situation examinée. Des données limitées sont parfois disponibles : pour certaines procédures, le détail des votes au niveau individuel n'est pas connu, mais les parlementaires qui ont voté contre leur groupe sont mentionnés (notamment en France), ce qui offre en soi assez d'élément pour spécifier l'*a priori* de manière adéquate.

des résultats sur de petits échantillons (les 15 membres du Conseil de sécurité de l'ONU, les 9 juges de la Cour suprême) il n'en va pas de même pour des parlements de plusieurs centaines de membres pour lesquels il peut arriver que l'on ne dispose effectivement *que* des données de vote agrégées, et de quelques éléments relatifs au mode de scrutin ou à la répartition des sièges entre les partis politiques. Le problème est assez similaire lorsqu'un grand nombre de covariables est considéré et que celles-ci peuvent se combiner (mesure de l'influence de clientèles électorales très différentes d'un-e représentant-e à l'autre, facteurs locaux pouvant intervenir dans certaines décisions et pour certains représentant-e-s, mais non pour d'autres, etc.). Outre le fait que ces situations affaiblissent, à nouveau, notre capacité à déterminer l'*a priori* elles rendent surtout très lente la convergence de l'algorithme. Etant donné que l'on ne simule qu'une seule composante à chaque itération, le temps nécessaire à la production de résultats est par conséquent beaucoup plus long.

Fort heureusement, des remèdes existent aux problèmes identifiés. Ils pourraient permettre d'adresser de manière frontale les difficultés propres au cadre Bayésien dans lequel se situe le modèle et à la lenteur de convergence de l'algorithme. De nombreuses méthodes ont ainsi été proposées dans la littérature pour les cas où les valeurs de départ ont été mal spécifiées, lorsque l'*a posteriori* est peu informatif ou en présence d'une chaîne de Markov rigide (Raftery et Lewis 1992) [6]. Pour qu'il s'applique au cas des parlements de manière optimale, une reconstruction de certaines composantes du modèle de Marbach (2017) devrait en conséquence être envisagée à l'aune de l'une ou l'autre de ces alternatives.

Lorsque l'on s'intéresse à l'activité parlementaire, il est donc possible d'étudier les déterminants du vote au niveau individuel quand on ne dispose que de résultats au niveau agrégé. Cet article a cherché à montrer qu'ignorer les données qui se présentaient sous cette forme ou suivre les recommandations des partisans de l'inférence écologique ne

6. Pour l'essentiel, ces alternatives se trouvent dans la famille des algorithmes de Metropolis-Hasting et, plus précisément, des algorithmes Monte-Carlo « Hamiltoniens » (Neal 2011).

constituaient pas des options pertinentes. On a ensuite vu que des travaux récents − dont le modèle de Marbach (2015, 2017) est selon nous la version la plus aboutie − pouvaient constituer un recours élégant pour établir les probabilités conditionnelles des votes au niveau individuel. Située dans un cadre Bayésien, les propriétés de ce modèle le rendent indéniablement plus difficile d'accès que les méthodes privilégiées par une grande partie la littérature. En raison de sa construction, il n'est pas non plus exempt de défauts qui ont été recensés et brièvement décrits. Si il requiert un certain nombre d'amendements que l'auteur de ces lignes s'efforce actuellement de lui apporter, nous considérons qu'il ouvre simultanément un certain nombre de perspectives méthodologiques fécondes, qui permettraient à terme aux chercheurs de s'affranchir de la contrainte posée par la non-publication de certaines données de vote au sein des organes parlementaires et plus généralement, dans la totalité des comités où l'accès aux données individuelles de décision est délicat, voire impossible.

Références

Anzia, S. et Moe, T. (2016) « Do Politicians Use Policy to Make Politics ? The Case of Public-Sector Labor Laws », *American Political Science Review*, 110(4) : 763-77.

Assemblée nationale (2017) « Les votes à l'Assemblée nationale », Site internet de l'Assemblée nationale, Fiche de synthèse n°44, consultée le 29 mai 2017.

Aurier, P. et Mejía, V. (2014) « Les modèles Logit et Probit multivariés pour la modélisation des achats simultanés : présentation, utilisation, intérêts et limites », *Recherche et Applications en Marketing*, 29(2) : 79-98.

Aydelotte, W. (1977) « Introduction », dans Aydelotte, W. (dir.) *The History of Parliamentary Behavior*, Princeton : Princeton University Press, 3-27.

Besag, J. (1974) « Spatial interaction and the statistical analysis of lattice systems », *Journal of the Royal Statistical Society*, 36(2) : 192–236.

Black, D. (1958) *The Theory of Committees and Elections*, London : Cambridge University Press.

Briatte, F. (2016) « Network patterns of legislative collaboration in twenty parliaments », *Network Science*, 4(2) : 266-71.

Broniatowski, D., Magee, C., Coughlin et al. (2009) 'Bayesian Analysis of Decision Making in Technical Expert Committees', *Proceedings of the 7th Annual Conference on Systems Engineering Research*, Loughborough University.

Carey, J. (2009) *Legislative Voting and Accountability*, New York : Cambridge University Press.

Carrubba, C. Gabel, M., Murrah, L. et al. (2006) « Off the Record : Unrecorded Legislative Votes, Selection Bias and Roll-Call Vote Analysis », *British Journal of Political Science*, 36 : 691-704.

De Fine Licht, J., Naurin, D., Esaiasson, P. et Gilljam, M. (2011) « Does transparency

generate legitimacy ? An experimental study of procedure acceptance of open and closed-door decision-making », *QoG Working Paper Series*, 8, September.

De Galembert, C., Rozenberg, O. et Vigour, C. (2014) *Faire parler le Parlement. Méthodes et enjeux de l'analyse des débats parlementaires pour les sciences sociales*, Paris : LGDJ.

Clinton, J. et Jackman, S. (2004) « The Statistical Analysis of Roll Call Data », *American Political Science Review*, 98(2) :355-370.

Finer, S. (1987) « Voting », in Bogdanor, V. (dir.), *The Blackwell Encyclopaedia of Political Science*, Oxford : Blackwell, 631.

Fry, M., D. Julius, L. Mahadeva, S. Roger, and G. Sterne (2000) « Key Issues in the Choice of Monetary Policy Framework », dans Mahadeva, L. et Sterne, G. (dir.), *Monetary Policy Frameworks in a Global Context*, London : Routledge.

Geyer, C. (2011) « Introduction to Markov Chain Monte Carlo », dans Gelman, A., Brooks, S. et Jones, G. et Meng, X. (dir.) *Handbook of Markov Chain Monte Carlo*, Boca Raton : CRC Press.

Gordon, S. et Bélanger, G. (1996) « Échantillonnage de Gibbs et autres applications économétriques des chaînes markoviennes », *L'Actualité économique*, 721 : 27-49.

Gourieroux, C. et Monfort, A. (1996) *Statistique et Modèles économétriques*, Paris : Economica.

Han, J-H. (2007) « Analysing Roll Calls of the European Parliament : a Bayesian Application », *European Union Politics*, 8(4) : 479-507.

Hix, S. (2002) « Parliamentary Behavior with Two Principals : Preferences, Parties and Voting in the European Parliament », *American Journal of Political Science*, 46(3) : 688-698.

Hug, S. (2009) « Selection Effects in Roll Call Votes », *British Journal of Political*

Science, 40(1) :225–235.

Hug, S. (2012) « Parliamentary Voting », Université de Genève, Working Paper, February.

Jackman, S. (2009) *Bayesian Analysis for the Social Sciences*, Chichester : Wiley.

Kaplan, D. et Depaoli, S. (2012) « Bayesian Structural Equation Modeling », dans Hoyle, R. (dir.) *Handbook of Structural Equation Modeling*, New York : Guilford Press, 650-673.

King, G. (1997) *A Solution to the Ecological Inference Problem : Reconstructing Individual Behavior from Aggregate Data*, Princeton : Princeton University Press.

King, G., Tanner, M. et Rosen, O. (1999) « Binomial-Beta Hierarchical Models for Ecological Inference », *Sociological Methods and Research* 28(1) : 61–90.

Lowell, L. (1901) « The Influence of Party Upon Legislation in England and America », *Annual Report of the American Historical Association*, 321-543.

Marbach, M. (2015) « A Discrete Choice Model for the Analysis of an Aggregate Voting Record », University of Mannheim, Working Paper, January 5.

Marbach, M. (2017) « Analyzing Decision Records from Committees », University of Mannheim, Working Paper, January 31.

Middlebrook, J. (2003) « Les méthodes de vote au sein des parlements », *Informations constitutionnelles et parlementaires*, 53(186) :42-67.

Neal, R. (2011) « MCMC Using Hamiltonian Dynamics », dans Gelman, A., Brooks, S. et Jones, G. et Meng, X. (dir.) *Handbook of Markov Chain Monte Carlo*, Boca Raton : CRC Press.

Raftery, A.E. et Lewis, S.M. (1992) « How many iterations in the Gibbs sampler ? » dans Bernardo, J. (dir.) *Bayesian Statistics*, New York : Oxford University Press, 763-773.

Robert, C. (2001) *The Bayesian Choice*, New York : Springer.

Robinson, W. (1950) « Ecological Correlations and the Behavior of Individuals », *American Sociological Review* 15(3) : 351–357.

Saalfeld, T. (1995) « On Dogs and Whips : Recorded Votes », dans Döring, H. (dir.) *Parliaments and Majority Rule in Western Europe*, New York : St. Martin's Press, 528–565.

Scharwz, D., Schädel, S. et Ladner, A. (2010) « Pre-Election Positions and Voting Behaviour in Parliament : Consistency among Swiss MPs », *Swiss Political Science Review*, 16(3) : 533-64.

Shugart, M. (1998) « The Inverse Relationship Between Party Strength and Executive Strength : A Theory of Politicians' Constitutional Choices », *British Journal of Political Science*, 28 : 1-29.

Squarcioni, L. (2016) *La dépendance au parti des députés. Conquérir, exercer et conserver son mandat au PS et à l'UMP*, Institut d'études politiques de Bordeaux, Thèse de doctorat (Science politique).

Tsebelis, G. (1995) « Decision Making in Political Systems : Veto Players in Presidentialism, Parliamentarism, Multicameralism and Multipartyism », *British Journal of Political Science*, 25(3) : 289-325.

VanDoren, P. (1990) « Can We Learn the Causes Of Congressional Decisions From Roll-Call Data ? », *Legislative Studies Quarterly*, 15(3) :311-40.

Zamora, S. (1980) « Voting in International Economic Organizations », *American Journal of International Law*, 74(3) : 566-608.