Quarante ans de délibération de la CNIL

Analyse thématique de l'évolution d'un corps de doctrine

THOMAS SOUBIRAN

CERAPS (UMR 8026 CNRS - Université de Lille)

Colloque Histoire, langues et textométrie

Paris, 17 janvier 2019

Introduction

- croissance des **collections** numérique et des **moyens de collecte** numérique depuis plusieurs décennies déjà
- nécessité de pouvoir **organiser** ces collections
- domaine de recherche **actif**
  - **différentes approches** ont été proposées
  - la présentation se focalisera sur une d'entre elle, **l'allocation Dirichlet latente** (LDA)
  - avec une application au **corpus des délibérations** de la Commission nationale informatique et libertés (CNIL)

Plan de la présentation

- La réglementation relative au traitement de données à caractère personnel
- Les modèles thématiques
- Analyse thématique des délibérations de la CNIL

La réglementation relative au traitement de données à données personnelles

La réglementation relatives aux DCP

- les réglementations sur les DCP ont commencé à apparaître dans les **années 70** du XX^{e} siècle
- en France, le traitement de DCP est encadré par la **n^{o} 78-17 du 6 janvier 1978** relative à l'informatique, aux fichiers et aux libertés (LIL)
- et le **règlement n^{o} 2016/679**, dit règlement européen sur la protection des données personnelles (RGPD)
- ces réglementations visent toute information permettant **d'identifier une personne physique** :

> *identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale*

La réglementation relatives aux DCP

- la particularité de la réglementation applicable aux DCP en France est qu'elle ne constitue qu'un **cadre général**
- qui s'articule autour de **cinq grands principes** :
  - **limitation de la finalité :** les données doivent être traitées de façon **compatible** avec une finalité **précise**
  - **minimisation des données :** seuls les informations **strictement nécessaires** à la réalisation de la finalité doivent être traités
  - **limitation de la conservation :** une fois la finalité réalisée, les informations doivent être **détruites** ou **anonymisées**
  - **information :** les personnes doivent être informées des traitements de données les concernant
  - **protection dès la conception (*privacy by design*) :** la protection des personnes et la sécurité des données doit être intégrée **dès la conception** du traitement

  L'information des personnes est le principe cardinal de la réglementation. Il renvoie au principe d'autodétermination informationnelle qui veut que les personnes doivent être en mesure de décider de l'utilisation des informations les concernant
- la réglementation ne propose donc qu'un cadre volontairement très **abstrait** qui peut donner lieu à de multiples interprétations lors de son application à des cas concrets
- qui confère une grande importance des **délibérations** de la CNIL dans l'application de la réglementation

La Cnil

- la Cnil est une **autorité administrative indépendante** crée par la loi n° 78-17 du 6 janvier 1978
- elle veille notamment à **la conformité** des traitements de DCP au regard de la réglementation en application
- la loi confère à la Commission un pouvoir **contrôle et de sanction administratives et pécuniaires**
- elle lui confère aussi un pouvoir **réglementaire**

  qui se manifeste par des normes simplifiées, des avis, des actes réglementaires uniques, des méthodologies de référence entre autres décisions prises en réunions plénières.

- ces différents textes forment **la doctrine** de la Cnil

Les délibérations de la Cnil

- les décisions de la Cnil, aussi appelées délibérations, permettent **d'établir la (non-)conformité** de traitements spécifiques
- les décisions constituent toutefois un corpus **vaste** (+18 000 documents) portant sur des traitements très **variés**
- d'où la nécessité **d'organiser** ce corpus pour l'explorer et l'interroger
- présentation d'un essai au moyen **d'un modèle thématique** (*Topic Model*), **l'allocation de Dirichlet latente** (LDA)

Les modèles thématiques

Les modèles thématiques

- le terme de modèles thématiques désigne un ensemble de modèles statistiques visant à **faire ressortir** des thèmes, des sujets ou encore des concepts de collections de documents
- pour **décrire, organiser et requérir ces collections**

  *ils ont en effet été développés dans une perspective de **récupération de l'information** et de **fouille textuelle***
- ils reposent sur le postulat que les documents ont une **structure inobservée** (latente)
- et que cette structure peut être inférée à partir des **co-occurrences** des mots contenus dans les documents
- le but est alors d'obtenir p. ex. **des classes ou des dimensions** permettant de faire ressortir la composition latente des documents

Les modèles thématiques

- différentes approches sont envisageables pour la caractérisation de ces espaces latents
- pour les besoins de l'exposé, on distinguera **deux types de méthodes** (Crain et al., 2012) :
  - **les méthodes de classification**
  - **les méthodes de réduction de la dimension des données**
- ainsi,
  - la classification utilise l'information sur la similarité des documents pour **les regrouper** de façon signifiante
  - l'appartenance des documents aux classes peut être peut être **unique ou multiple**
  - la réduction de dimensions cherche plutôt à faire ressortir **des traits saillants** des documents comme des thèmes.

Les modèles thématiques

- la classification **s'apparente** à une représentation en dimension réduite
- toutefois les classes ne représentent pas nécessairement un trait, la similarité entre documents **en combinant généralement plusieurs**
- les modèles thématiques comme la `LDA` **combinent les deux approches**
- dans le sens où les documents peuvent appartenir **à plusieurs classes** qui correspondent à différents traits apparaissant dans le corpus

  chaque documents se voit assigner différents poids mesurant la force de l'appartenance à une classe ainsi que leur position dans l'espace réduit des dimensions.

- dans ce qui suit, l'espace latent que l'on souhaite faire ressortir est celui des **catégories de traitement**

  soit des classes liant données, opérations sur ces données et finalité de ces opérations

Le sens des classes

- quelle que soit la méthode utilisée, si l'objet est dégager des thèmes, les classes peuvent manifester **d'autres caractéristiques** des textes comme des registres ou des tournures
- comme on le verra, certaines classes peuvent caractériser des **vocabulaires propres** aux délibérations en général
- et ne pas manifester une catégorie de traitement en particulier
- de ce point de vue les traits constitutifs de l'espace latent dépendent aussi fortement de **la préparation du corpus**
- d'autres classes peuvent relever **d'un bruit de fond** sans réellement faire sens, la co-occurrence n'étant pas toujours sémantiquement informative

Exemple : les textes issus du web comme les différents type de post contiennent souvent des citations d'autres post ce qui conduit à gonfler le nombres des co-occurrences sans pour autant transmettre plus d'information sur les thèmes

Factorisations

- les modèles thématiques comme `LDA` (et d'autres) reposent sur l'analyse de matrice **termes–documents** :

$$\begin{pmatrix} n_{1,1} & \dots & n_{1,V} \\ \vdots & \ddots & \vdots \\ n_{1,D} & \dots & n_{D,V} \end{pmatrix} \tag{1}$$

  où $n_{d,v}$ compte le nombre d'occurrences du terme v dans le document d

- sous une hypothèse **d'interchangeabilité** des mots (*bag-of-words*)

  l'ordre d'apparition et la succession des mots n'a pas d'importance

- les modèles thématiques visent à résumer l'information contenue dans la matrice de co-occurrence en deux matrices de dimensions réduites projetant les documents et les termes dans **l'espace des traits**
- elle partage ce type de **factorisation** avec d'autres méthodes

  dont l'objet ne se limite pas aux données textuelles

- la différence est qu'ici, la factorisation est obtenue en se fondant sur **un modèle probabiliste explicite** reposant sur un modèle génératif décrivant le processus générant les documents

Factorisations

- une méthode courante utilisée pour **la réduction de dimensionnalité** est la **décomposition en valeur singulière**

  qui généralise la décomposition en valeurs propres à des matrices non-symétriques

- elle consiste à factoriser la matrice de départ **en trois matrices** :

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\intercal \tag{2}$$

$$= \underset{D\times K}{(\mathbf{u}_1, \ldots, \mathbf{u}_K)} \begin{pmatrix} \sigma & & \\ & \ddots & \\ & & \sigma_K \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^\intercal \\ \vdots \\ \mathbf{v}_K^\intercal \end{pmatrix}_{K\times V} \tag{3}$$

$$= \sum_k \sigma_k \mathbf{u}_k \mathbf{v}_k^\intercal \tag{4}$$

Factorisations

- la factorisation de la matrice en matrices de rang moindre se retrouve p. ex. dans **l'ACP** ou **l'AFC**
- ou **la LSI** qui est un des premier modèle thématique a avoir été proposé
- où il s'agit de **projeter** les lignes et les colonnes dans un espace latent

  cet espace étant constitué par les semi-axes de l'hyper-ellipse qui enveloppe tous les vecteurs de la matrice

- **Exemple : l'AFC**
  - le problème de l'AFC est trouver la solution **au système d'équation** suivant :

    $$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{U}\Sigma$$

    $$\mathbf{G} = \mathbf{D}_c^{-1}\mathbf{V}\Sigma$$

    où $\mathbf{F}$ et $\mathbf{G}$ sont respectivement les coordonnées des lignes et des colonnes dans le nouveau repère orthonormé
  - qui correspond à **la décomposition en valeurs singulières**

    $$(\mathbf{Z} - \mathbf{r}\mathbf{c}^\intercal) = \mathbf{U}\Sigma\mathbf{V}^\intercal$$

-

Factorisations

- dans les faits, ces méthodes diffèrent principalement par **la matrice factorisée**
- l'`AFC` utilise **la distance du χ^2**
- La `LSI` pondère les entrées de la matrice de co-occurrence en utilisant **des pondérations de type** `tf-idf` (*term frequency – inverse document frequency*)

$$\texttt{td} - \texttt{idf} = \texttt{td} \times \texttt{idf} \tag{5}$$

`td` :

$$\begin{cases} 1 & \text{si } w_n \in \mathbf{w}_d \\ 0 & \text{sinon} \end{cases}$$

$w_{d,n}$ (# d'occurrences dans le document)

$w_{d,n}/w_{d,+}$ (fréquence relative)

`idf` :

$$\log\left(\frac{D}{\texttt{df}}\right)$$

$$\log\left(\frac{D}{1+\texttt{df}}\right)$$

`df` désigne la fréquence des documents, soit le nombre de documents dans lesquels le mot w_n apparaît.

- la pondération vise à faire ressortir **les mots les plus fréquents** dans chacun des documents tout en pénalisant ceux qui apparaissent **dans de nombreux documents**
- les matrices **U** et **V** donnant **les coordonnées** des documents et les dermes du vocabulaire dans **un espace (sémantique) latent**

Factorisations

- la LDA cherche aussi à produire **deux matrices** ayant une interprétation géométrique similaire
  - proportion de chaque classe dans chaque document

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix} = \begin{pmatrix} \theta_{1,1} & \dots & \theta_{1,K} \\ \vdots & \ddots & \vdots \\ \theta_{1,D} & \dots & \theta_{D,K} \end{pmatrix} \quad \text{avec} \quad \sum_{k=1}^{k} \phi_{d,k} = 1 \tag{6}$$

  - proportion de chaque terme dans chaque classe

$$\phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_K \end{pmatrix} = \begin{pmatrix} \phi_{1,1} & \dots & \phi_{1,V} \\ \vdots & \ddots & \vdots \\ \phi_{1,K} & \dots & \phi_{K,V} \end{pmatrix} \quad \text{avec} \quad \sum_{v=1}^{V} \phi_{k,v} = 1 \tag{7}$$

- ces deux matrices représentant, là aussi, les positions des documents et du vocabulaire dans **un espace sémantique latent**

  cette fois, non pas dans un repère orthonormé mais un simplex de probabilités

- la LDA procède toutefois de façon très différente
- en se basant sur **un modèle génératif** explicite des documents

Le modèle génératif de la LDA

- le modèle génératif de LDA repose sur l'intuition que chaque document est constitué de **plusieurs thèmes**
- ces thèmes apparaissent dans chaque document dans dans **des proportions différentes**
- chaque mot relève **d'un (et un seul) des thèmes** propres à chaque document
- une fois le thème du mot déterminé, le mot est **sélectionné** parmi les termes du thème retenu

Le modèle génératif de la LDA

1. pour $k = 1, \ldots, K$:
   1.a $\phi^{(k)} \sim Dir(\beta)$
2. pour tous les documents $d \in \mathcal{D}$:
   2.a $\theta_d \sim Dir(\alpha)$
   2.b pour tous les mots $w_{d,n}$ du document :
      2.b.i $z_{d,n} \sim Cat(\theta_d)$
      2.b.ii $w_{d,n} \sim Cat(\phi_{z_{d,n}})$

- globalement,
  - tirage des probabilités **des termes** de chaque thèmes
- puis, pour chaque document,
  - tirage des probabilités **des thèmes** pour chaque document
  - puis, pour chaque position (indice) de mots de chaque document,
    - tirage **du thème de chaque mot** : le thème z_{dn} de chaque mot du document est tiré de la distribution des thèmes du document
    - tirage **du mot** : pour chaque indice du document, un mot est tiré de la distribution multinomiale associée au thème retenu.

L'allocation Dirichlet latente

- le modèle génératif de la LDA conduit à l'expression suivante de **la probabilité jointe**

$$p(\mathbf{W}, \mathbf{Z}, \theta, \phi | \alpha, \beta) = \prod_{k}^{K} p(\phi_k | \beta) \prod_{d}^{D} p(\theta_d | \alpha) \prod_{n}^{n_j} p(z_{d,n} | \theta_d) \, p(w_{d,n} | \phi_{z_{d,n}}) \quad (8)$$

- les proportions des thèmes dans chaque document ne sont **pas considérées** comme des paramètres
- ce cas de figure correspond à un autre modèle probabiliste, **l'indexation sémantique latente probabiliste** (pLSI)

$$p(w, d) = p(d) \sum_{k}^{K} p(w | k) \, p(k | d) \quad (9)$$

- à la place, la LDA adopte une approche **bayésienne**
- où les proportions des thèmes θ_d dans les documents sont traitées comme **des variables latentes** paramétrées par une distribution (β)
- de même que **la distribution des mots** ϕ_k

L'allocation Dirichlet latente

- traiter les distributions des poids et des mots comme des variables latentes **complique l'estimation du modèle**
- les proportions doivent en effet **être intégrées**

$$p(\boldsymbol{W}, |\alpha, \beta) = \int_{\phi} \int_{\theta} p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta) d\theta d\phi \tag{10}$$

- d'où l'utilisation d'un **a priori conjugué** à la distribution multinomiale sur les $\boldsymbol{\theta}_d$ et $\boldsymbol{\phi}_k$: la distribution de Dirichlet

$$p(\mathbf{x}|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \dots x_k^{\alpha_K - 1} \tag{11}$$

- $\boldsymbol{\theta}_d$ et $\boldsymbol{\phi}_k$ sont inférées **après l'estimation**

L'allocation Dirichlet latente

- (Blei, Ng et Jordan, 2003) motive cette approche **par contraste** à la pLSI
- le modèle pLSI nécessite **l'estimation des probabilités** $p(z|d)$ des thèmes pour chaque document

  $p(d) = 0$ *pour un document en dehors du corpus*

- ces probabilités sont donc propres à un corpus et les résultats y sont **contingentés**

  avec un risque de surapprentissage

- la taille $K \times D$ de $p(z|d)$ à estimer **croît linéairement** avec D ce qui comporte un risque de sur-ajustement (ou de sur-apprentissage)

Analyse thématique des délibérations de la CNIL

Les délibérations de la CNIL

- **18 551 décisions** rendues entre 1979 et 2017
- disponibles sur la **page** de la CNIL du site data.gouv.fr
- chaque document comporte :
  - un identifiant
  - la date de la délibération
  - le titre de la délibération
  - le texte de la délibération
  - la nature de la délibération

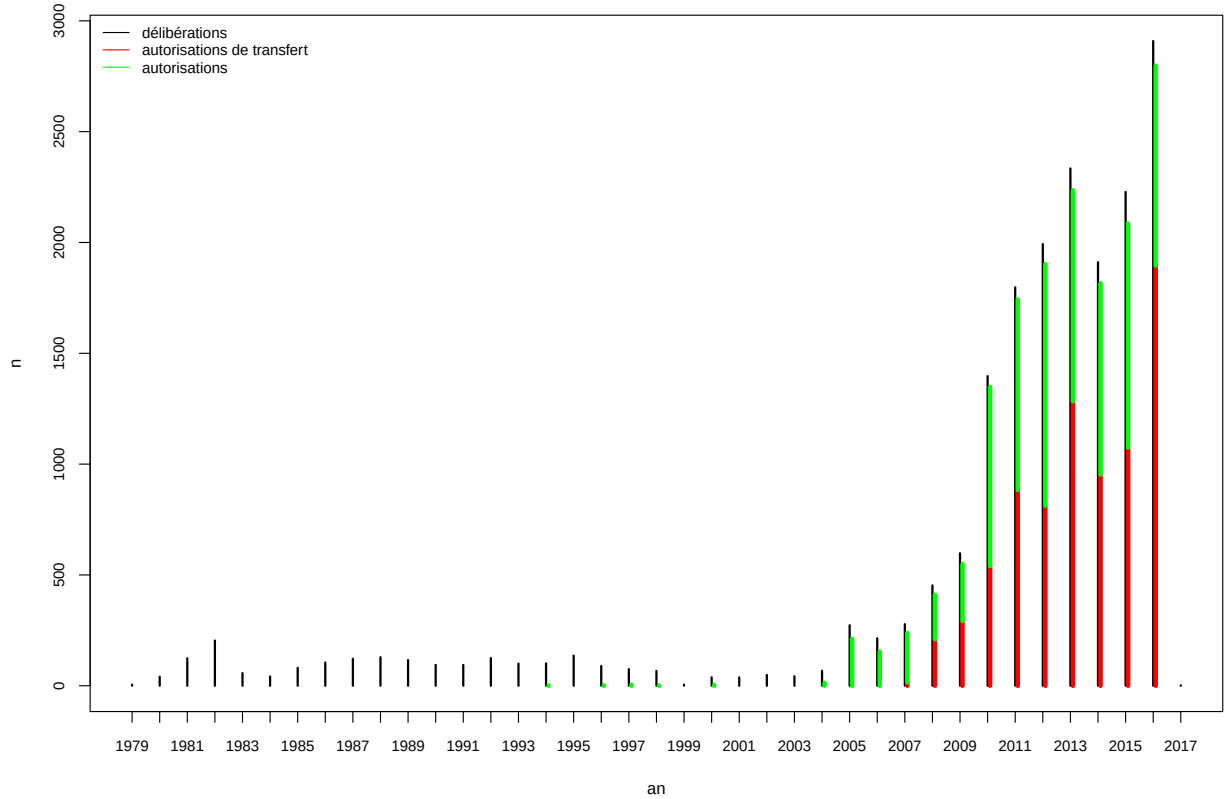
    Note : ce champ ne semble pas avoir fait l'objet d'une standardisation (nombreuses variations y compris dans la graphie) et le niveau de renseignement fourni peut varier d'une occurrence à l'autre. Ainsi, certaines demandes d'autorisation de transfert sont classées comme demandes d'autorisation

Les délibérations de la Cnil

- le nombre annuel de délibérations a d'abord été relativement **stable** (graphique 1)
- puis a progressivement **décru** à partir de la seconde moitié des années 1990
- avant de **croître** à nouveau et plus fortement après 2004

  l'année 2004 marque en effet la traduction dans le droit français de la directive européenne 95/46/CE *sur la protection des données*

- **Note :** les proportions apparaissant sur le graphique sont sous-estimées car des délibérations classées comme des autorisations (sans plus de précisions) concernent des autorisation de transfert



Grahique 1: Nombre de délibérations de la CNIL par an

Préparation du corpus

- analyse **morpho-syntaxique** avec `TreeTagger`
- ont été **conservés** les verbes, les noms, les adjectifs, les abréviations ainsi que les occurrences n'ayant pas pu être étiquetées

  en effet, l'analyseur syntaxique a connu quelques ratés, notamment du fait de l'absence d'informations sur le formatage des documents originaux comme les alinéas
- les mots les plus rares ont ensuite été supprimés du corpus
- sans pour autant chercher à **trop réduire** le vocabulaire
- en effet, les délibérations portent généralement sur des sujets précis et peuvent employer un vocabulaire **très spécifique** qui peut ne concerner que quelques délibérations
- un émondage trop drastique risquait de faire **disparaître** des catégories de traitement rares

  (Chuang et al., 2015)
- n'ont donc été conservés que les mots apparaissant dans **au moins trois délibérations**

Préparation du corpus

- les autorisations de transfert constituent **une catégorie à part** de délibération

  elles se caractérisent par des textes courts ne faisant mention que du responsable de traitement, des données concernées et de leur destination

- elles ont donc été **exclues** du corpus car elles ne recèlent que peu d'informations sur la doctrine de la CNIL
- au final, l'analyse porte donc sur **10 297 délibérations**, soit 55% du total et sur un vocabulaire comportant **10 639 termes**

Description des classes

- présentation des résultats d'un modèle à **120 classes**

  Note : la significativité statistique n'est pas la significativité sémantique. . .

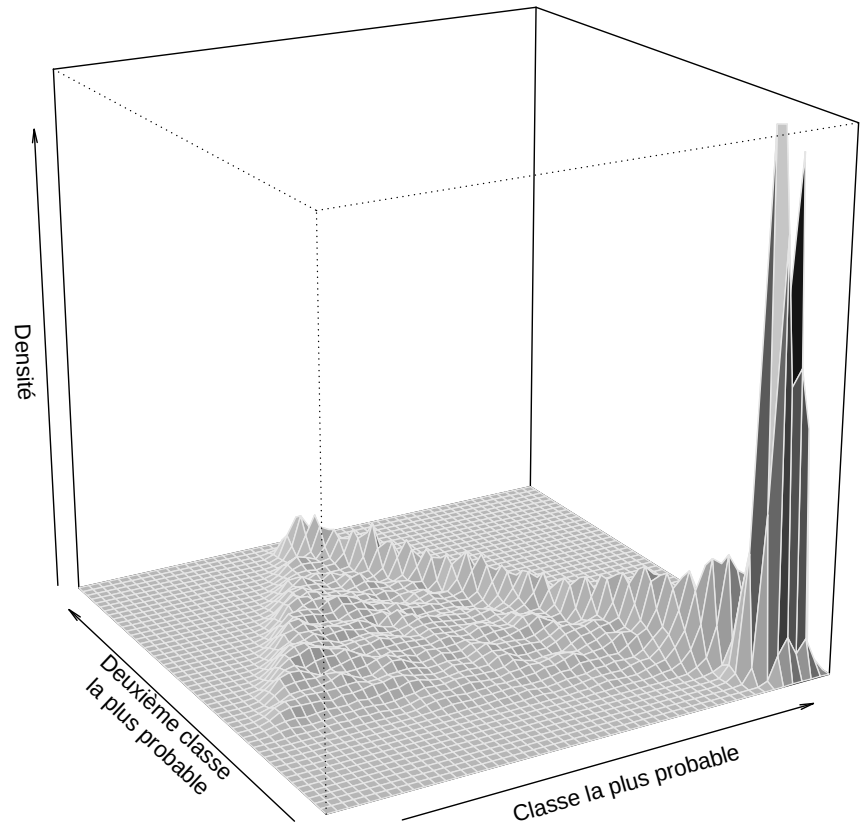
- nombre suffisant pour permettre une première **exploration** du corpus
- même si un nombre **supérieur** de classes semble nécessaire pour pleinement rendre compte des composantes du corpus

  du point de vue des mesures d'ajustement utilisées et de l'inspection des classes

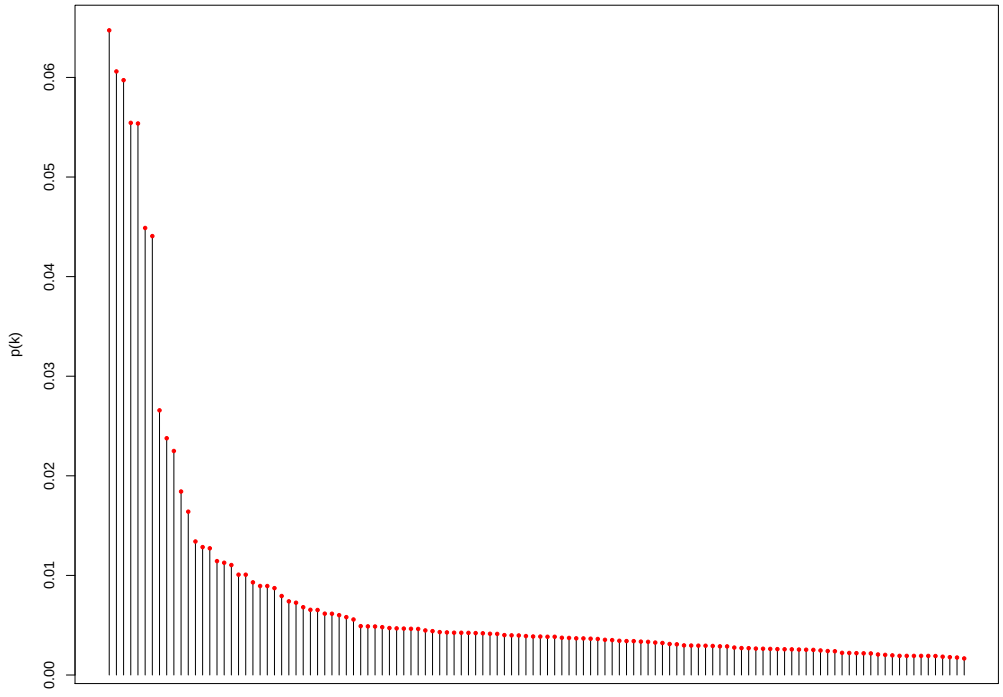
- globalement, les délibérations apparaissent :
  - **mono-thématiques** (graphique 2)

    ce qui renvoie au modalités propres de l'application de la réglementation relatives aux DCP : la réglementation ne définit que des principes généraux et la conformité de chaque traitement doit être établie au cas par cas.

  - **très diversifiées** (graphique 3)



Grahique 2: Densité des probabilités postérieures des deux classes les plus probables des documents



Grahique 3: Probabilités d'observer les classes dans les documents

Description des classes

- on peut toutefois noter une prévalence des traitements relatifs **au domaine de la santé**
- 8 des 10 classes les plus probables concernent des **autorisations** dans le domaine de la recherche médicale
- 11 des 20 classes les plus importantes si on y adjoint le **domaine médico-social**
- la prévalence du domaine médical peut aussi être attestée par le **nuage de mots** des mots les plus saillants dans le corpus au regard du modèle (graphique 4)

Grahique 4: Nuage des mots les plus saillants dans le corpus des délibérations de la CNIL

Description des classes

Parmi les 20 premières classes, on trouve aussi :

- des demandes d'avis provenant de **collectivités locales** concernant différents traitements de DCP portant sur leurs administrés (élections, foncier, fiscalité,. . .). La classe se caractérise par les lemmes suivants : `impot`, `habitation`, `taxe`, `ccid`, `9151`, `fiscal`, `prudhomme`, `municipal`, `mairie`, `commune`, `foncier`, `electeur`, `electoral`
- des demandes d'autorisation de mise en œuvre de **dispositifs d'alerte professionnelle**. La classe se caractérise par les lemmes suivants : `alerte`, `au004`, `corruption`, `anticoncurrentielles`, `signalement`, `dispositif`, `ethique`, `discrimination`
- des demandes d'autorisation de mise en œuvre de dispositifs **biométriques**. La classe se caractérise par les lemmes suivants : `gabarit`, `veineux`, `doigt`, `biometrique`, `reconnaissance`, `lecteur`, `empreinte`
- des demandes **d'autorisation de transfert** de données principalement dans le domaine bancaire
- ainsi que trois classes correspondant aux délibérations en elles-mêmes (cf. *infra*)

Note : Les mots ont été sélectionnés en utilisant le score suivant :

$$\text{term-score}_{k,v} = \hat{\phi}_{k,v} \log \left[\hat{\phi}_{k,v} \Big/ \left(\prod_{k'=1}^{K} \hat{\phi}_{k',v} \right)^{1/K} \right]$$

Description des classes

De façon moins marquée, d'autre types de traitements transparaissent aussi dans le nuage de mots :

- les traitements de la **statistique publique**. L'INSEE tient une part prépondérant dans ces différentes classes, principalement pour ce qui relève de l'organisation des recensements et la réalisation de ses enquêtes auprès des ménages. Les enquêtes de la statistique publique relatives au domaine de la santé ou du médico-social comme celles diligentées par la DRESS occupent aussi une part non négligeable de ces délibérations.
- les demandes d'avis ou d'autorisation émanant du **ministère de l'intérieur** ou du **ministère de la Justice** pour la mise en place de différents fichiers de police (empreintes, infractions, antécédents judiciaires, auteurs d'infractions terroristes), dans le cadre de l'activité des tribunaux de justice ou relatives au détenus
- la lutte contre la **fraude** (bancaire ou aux allocations), le blanchissement d'argent (là encore, en relation avec une entreprise terroriste dans des délibérations plus récentes)

Note : le modèle a aussi identifié des catégories de traitement ne se manifestant que dans une nombre très limité de délibération

DRM, *réparations des victimes des persécutions antisémites pendant le régime de Vichy,. . .*

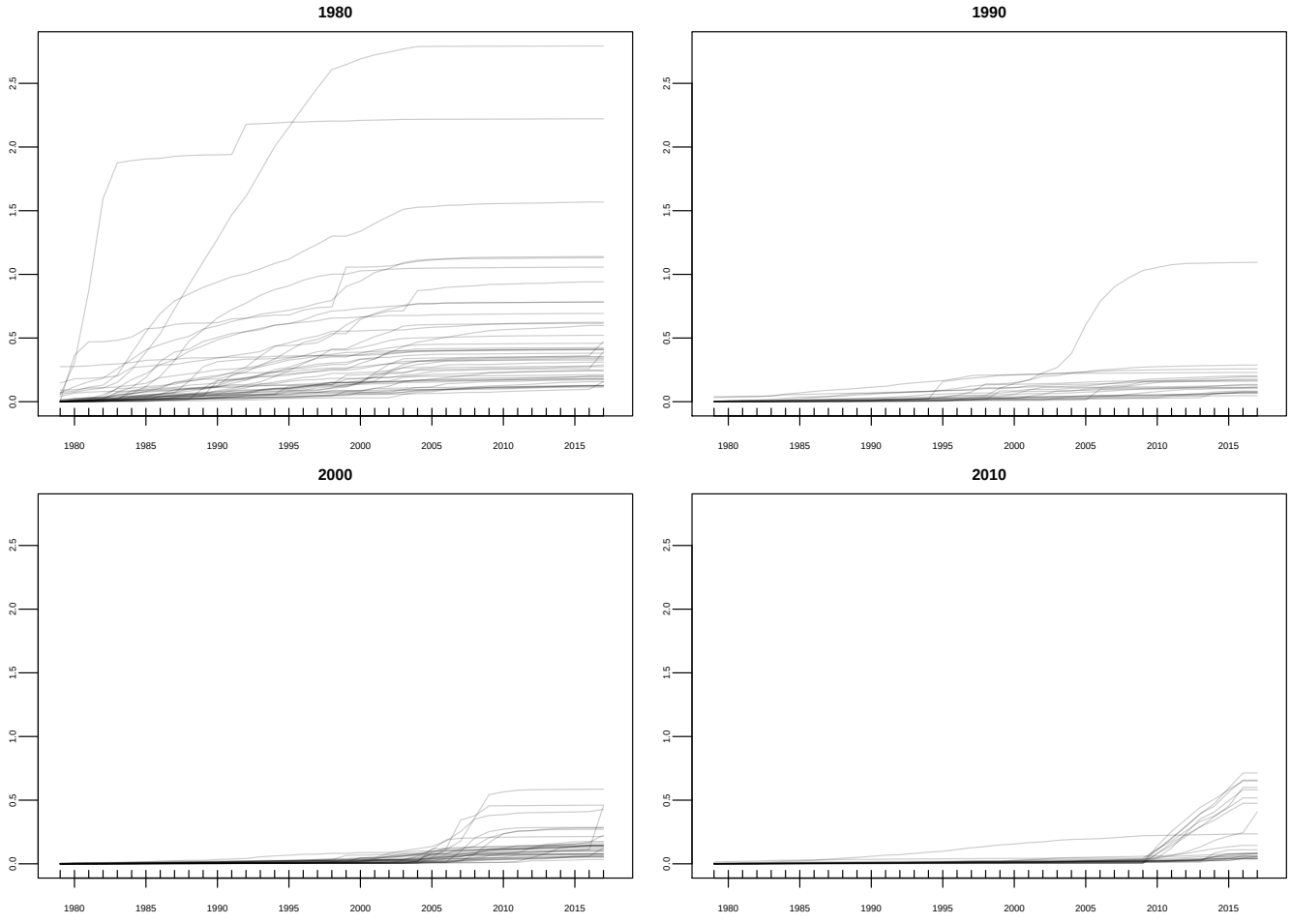
Évolution dans le temps

- les probabilités d'apparitions des classes ne sont **pas constantes** dans le temps (graphique (5))

  Le graphique reprend l'approche utilisée par (Griffiths et M. Steyvers, 2004) pour détecter les tendances dans les thèmes. Elle consiste à calculer la moyenne des $\theta_{d,k}$ par an pour chaque classe. Le graphique projette les valeurs cumulées afin de mieux faire ressortir les tendances.

- on notera tout d'abord là aussi la rupture au milieu des années 2000 qui correspond à la **transposition de la directive** 95/46/CE dans le droit français. .
- ainsi, les probabilités moyennes cumulées de nombreuses classes **cessent de croître** aux alentours de cette date
- de plus, **les catégories de traitements** les plus saillantes varient dans le temps
- ainsi que **les responsables de traitement**

  l'importance des traitements issus du secteur public tend à s'estomper avec le temps (cf. « l'affaire Safari »)



Grahique 5: Évolution des probabilités moyennes d'observer les classes dans les documents

Évolution dans le temps

- les délibérations **des années 1980** concernent, entre autres :
  - les traitements de DCP par les collectivités locales
  - le traitement de DCP par les entreprises
  - l'organisation du système de santé
  - l'utilisation du NIR
  - les prestations sociales
  - le traitement de DCP par les organisme d'assurances et les mutuelles
  - ou encore l'organisation du recensement par l'INSEE.
- celles **des années 1990** concernent, elles
  - l'assurance maladie
  - l'accompagnement social
  - mais aussi internet (« réseau international ouvert dénommé "internet" »)

    Note : leur nombre croît surtout dans les deux décennies suivantes

Évolution dans le temps

- **les années 2000** voient notamment le développement des délibération portant sur
  - la biométrie
  - la coordination des soins
  - la mise en place de systèmes d'alerte professionnelle
  - la publication de normes simplifiées et d'autorisations uniques
  - la mise en place de systèmes d'information géographique (cadastre, surveillance de flottes de véhicules)
  - les enquêtes de la statistique publique
  - mais aussi le prononcé de sanctions.
- **les années 2010** marquent le développement des délibérations relatives à la recherche médicale. On notera aussi les demandes de labellisation ainsi que les sanctions à l'encontre d'entreprises comme Google.

Évaluation des classes

- les classes résultant de l'analyse ne sont **pas toujours interprétables** comme des thèmes
- (AlSumait et al., 2009) ont proposé un score pour distinguer **les classes insignifiantes** (« *junk topic* ») des autres
- ce score est une combinaison pondérée de **trois mesures de dissimilarité** :

$$\mathcal{S}^{(k)} = \sum_{u}^{U} \Psi_u \, \mathcal{S}_u^{(k)} \tag{12}$$

avec $\mathcal{S}_u^{(k)}$, la mesure, Ψ_u, son poids total et U, le nombre de mesures

Mesures de dissimilarités

- une mesure **d'uniformité** qui mesure la dissimilarités entre $\phi^{(k)}$ et la distribution

$$p(w_n|\Omega^{\mathcal{U}}) = \frac{1}{V}, \ \forall\, n \in \{1, \ldots, V\} \tag{13}$$

où tous les mots du vocabulaire sont considérés comme équiprobables.

- une mesure **de vacuité** qui mesure la dissimilarité entre $\phi^{(k)}$ et la distribution marginale des mots prédite par le modèle

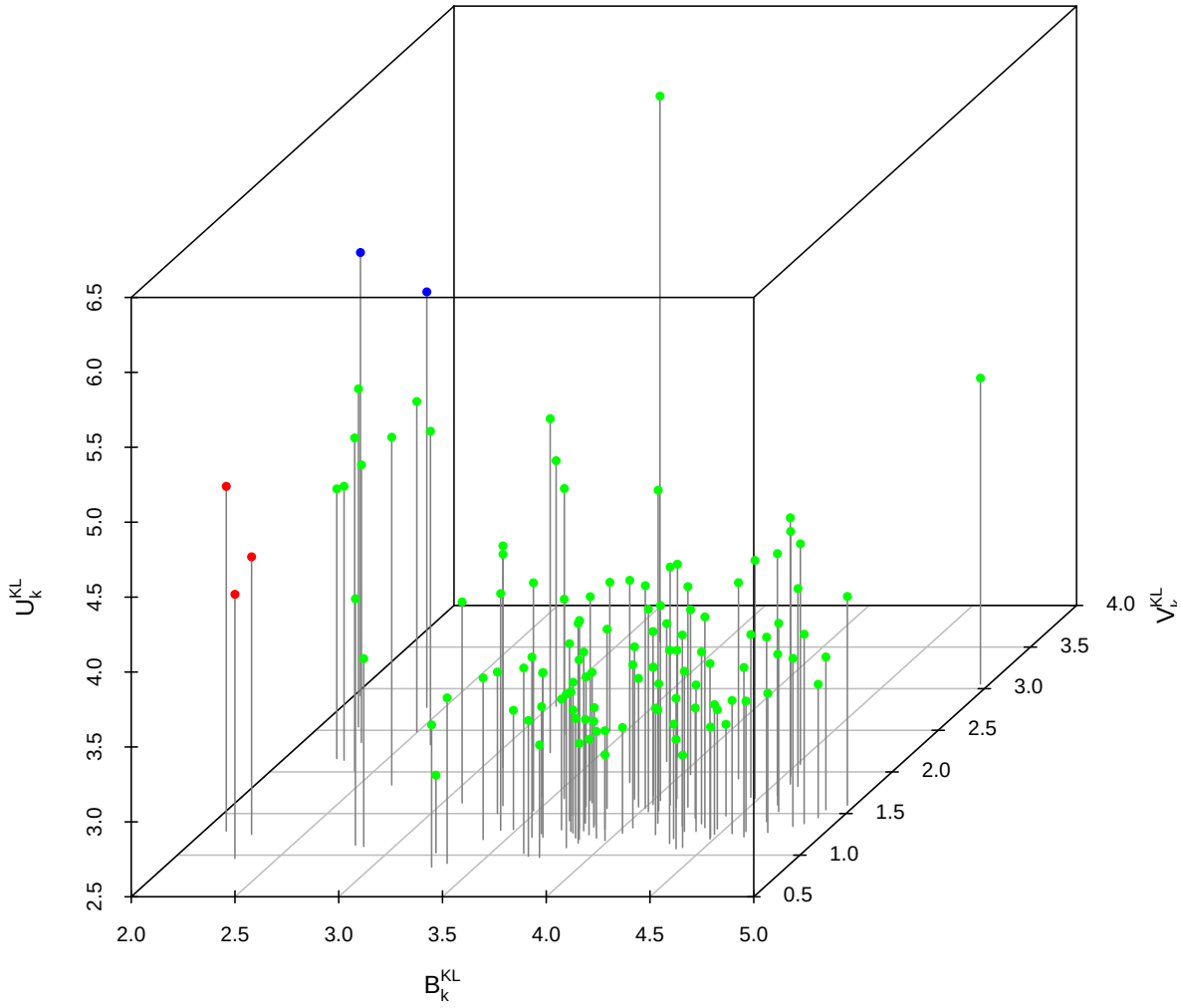
$$p(w_n|\Omega^{\mathcal{V}}) = \sum_{k=1}^{K} p(w_n|k)\, p(k) \tag{14}$$

- une mesure **de communalité** des classes qui mesure la dissimilarité cette fois entre $\theta^{(k)}$ et la distribution

$$p(d_m|\Omega^{\mathcal{B}}) = \frac{1}{M}, \ \forall\, d \in \{1, \ldots, D\} \tag{15}$$

qui donne une mesure de la dispersion des thèmes dans le corpus

- ces dissimilarités peuvent, entre autres possibilités, être calculées avec la divergence de Kullback-Leibler.



Grahique 6: Dissimilarités $\mathcal{U}_k^{KL}$, $\mathcal{V}_k^{KL}$ et $\mathcal{B}_k^{KL}$

Distributions spécifiques

- **trois classes** (en rouge) se retrouvent côte-à-côte en étant à la fois (graphique 6) :
  - très répandue parmi les délibérations ($D_{KL}(\theta^{(k)}, \Omega^{\mathcal{B}})$)
  - tout en en présentant un vocabulaire spécifique ($D_{KL}(\phi^{(k)}, \Omega^{\mathcal{U}})$)
  - et que l'on retrouve dans de nombreux documents ($D_{KL}(\phi^{(k)}, \Omega^{\mathcal{V}})$)
- ces classes renvoient au **vocabulaire propres** des délibérations
- ces trois classes sont plus particulièrement marquées par **les noms** figurant dans les délibérations
- ainsi, parmi les mots les plus saillants de ces classes, on trouve les noms des **trois présidents** ayant rendu le plus de décisions :
  - l'ancien directeur du quotidien *Le Monde* Jacques Fauvet qui a présidé la CNIL de 1984 à 1999
  - le sénateur du Nord Alex Türk qui a présidé la CNIL de 2002 à 2011
  - la conseillère d'État Isabelle Falque-Pierrotin qui préside la commission depuis 2011
- ces classes sont donc aussi caractéristiques **des périodes** auxquelles les décisions ont été rendues

Autres classes

- **deux classes** (en rouge) plus éloignées sur l'axe ($D_{KL}(\phi^{(k)}, \Omega^{\mathcal{V}})$) ne constituent pas non plus des catégories de traitements de DCP
- la première fait plutôt référence aux données traitées et, particulièrement, aux **données directement identifiantes** (`prenom`, `nom`, `date`, `naissance`, `adresse`, `numero`, `telephone`, `sexe`, ...)
- la seconde fait plutôt référence aux traitements réalisés dans des **administrations publiques** (`direction`, `departemental`, `general`, `affaire`, `agent`, `service`, ...)
- **Note :** les classes voisines concernent des autorisations dans le domaine de la recherche médicale

Extensions et variantes de la LDA

- l'examen des classes montre que le modèle LDA original **ne prend pas en compte** certaines caractéristiques des textes
- c'est pourquoi, de nombreuses **variantes et extensions** ont été proposées :
  - **vocabulaire propre** au corpus et au documents (Chemudugunta, Smyth et Mark Steyvers, 2007)
  - **liens entre les classes** en ajoutant une structure de corrélation (Lafferty et Blei, 2006) ou une structure hiérarchique (Blei, Jordan et al., 2003)
  - **dimension temporelle** : évolution de la prévalence des thèmes et des thèmes eux-mêmes (Blei et Lafferty, 2006)

Conclusion

Conclusion

- l'application de la LDA aux délibérations donne des résultats **qualitativement pertinents**

  pas de classes incompréhensibles

- pour **l'exploration de corpus**
- et permet de dégager **d'autres aspects** des textes que les thèmes par des analyses descriptives

  comme le vocabulaire spécifique ou l'évolution des thèmes

- la LDA procure de plus **un cadre très flexible**
- qui autorise **de nombreuses extensions** pour mettre en œuvre des mdoèles génératifs plus complexes

Merci pour votre attention

Bibliographie

Bibliographie I

AlSumait, Loulwah, Daniel Barbará, James Gentle et Carlotta Domeniconi (2009), « Topic Significance Ranking of LDA Generative Models », *Machine Learning and Knowledge Discovery in Databases*. Sous la dir. de Wray Buntine, Marko Grobelnik, Dunja Mladenić et John Shawe-Taylor, Berlin, Heidelberg, Springer Berlin Heidelberg, p. 67–82.

Blei, David M., Michael I. Jordan, Thomas L. Griffiths et Joshua B. Tenenbaum (2003), « Hierarchical Topic Models and the Nested Chinese Restaurant Process », *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, Whistler, British Columbia, Canada, MIT Press, p. 17–24.

Blei, David M. et John D. Lafferty (2006), « Dynamic Topic Models », *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, Pittsburgh, Pennsylvania, USA, ACM, p. 113–120. doi : 10.1145/1143844.1143859. url : http://doi.acm.org/10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng et Michael I. Jordan (2003), « Latent Dirichlet Allocation », *J. Mach. Learn. Res.* 3, p. 993–1022.

Chemudugunta, Chaitanya, Padhraic Smyth et Mark Steyvers (2007), "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model". *Advances in Neural Information Processing Systems 19*. Sous la dir. de B. Schölkopf, J. C. Platt et T. Hoffman, MIT Press, p. 241–248. url : http://papers.nips.cc/paper/2994-modeling-general-and-specific-aspects-of-documents-with-a-probabilistic-topic-model.pdf.

Bibliographie II

TopicCheck : Interactive Alignment for Assessing Topic Model Stability (2015), Denver, Colorado,

Crain, Steven P., Ke Zhou, Shuang-Hong Yang et Hongyuan Zha (2012), "Dimensionality Reduction and Topic Modeling : From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond". *Mining Text Data*. Sous la dir. de Charu C. Aggarwal et ChengXiang Zhai, Boston, MA, Springer US, p. 129–161. doi : 10.1007/978-1-4614-3223-4_5. url : https://doi.org/10.1007/978-1-4614-3223-4_5.

Griffiths, T. L. et M. Steyvers (2004), « Finding scientific topics », *Proceedings of the National Academy of Sciences*, Suppl. 1, 101, p. 5228–5235.

Lafferty, John D. et David M. Blei (2006), "Correlated Topic Models". *Advances in Neural Information Processing Systems 18*. Sous la dir. d'Y. Weiss, B. Schölkopf et J. C. Platt, MIT Press, p. 147–154. url : http://papers.nips.cc/paper/2906-correlated-topic-models.pdf.