Modéliser l'incertitude : une approche bayésienne des données textuelles

Guilherme D. Garcia

Université Laval
Centre for Research on Brain, Language and Music (CRBLM)

Linguistique, traduction et humanités numériques à l'ère de l'intelligence artificielle

25–26 septembre, 2025 • UQAC

UNIVERSITÉ LAVAL

CRBLM
Centre for Research on Brain, Language and Music

Aujourd'hui

1. L'**analyse de données** à partir de la statistique bayésienne
2. L'incertitude autour des estimation des modèles
3. Exemple illustratif : la variation des sentiments dans un roman

L'IA et l'analyse de texte

Nouvelles possibilités, nouveaux défis

L'essor des outils d'IA pour le texte :

- Grands modèles de langage (ChatGPT, Claude...)
- Exemple : analyse de sentiment automatisée
- Extraction d'information à grande échelle

Mais... Comment évaluer la fiabilité ?

- Résultats opaques (*boîtes noires*)
- Fausse précision (« 73,2 % de sentiment positif »)
- Biais cachés dans les données d'entraînement

Une approche alternative

L'analyse bayésienne

☞ Comment analyser les textes de manière **transparente** et **rigoureuse** face à l'incertitude ?

Approche traditionnelle :

- Estimations ponctuelles
- Tests de significativité
- Résultats binaires (significatif/non)

Approche bayésienne :

- Quantification de l'incertitude
- Intégration de connaissances préalables
- Énoncés probabilistes directs

☞ **Objectif :** Démontrer quelques avantages de l'analyse bayésienne pour les humanités numériques avec *Madame Bovary* comme cas d'étude

Les fondements de l'approche bayésienne

Le théorème de Bayes

$$P(\text{hypothèse} \mid \text{données}) = \frac{P(\text{données} \mid \text{hypothèse}) \times P(\text{hypothèse})}{P(\text{données})} \tag{1}$$

En termes pratiques :

$$\text{Ce qu'on croit} \propto \text{Ce qu'on observe} \times \text{Ce qu'on savait avant} \tag{2}$$

Philosophie bayésienne

☞ L'incertitude est intrinsèque à la connaissance. On met à jour nos croyances de manière rationnelle à mesure qu'on reçoit de nouvelles données. La logique bayésienne de façon visuelle.

Les fondements de l'approche bayésienne

Uns illustration de deux paramètres

- On ne peut pas calculer la loi a posteriori elle-même (impossible en pratique)

☞ Mais on peut l'approximer en tirant des échantillons de cette loi

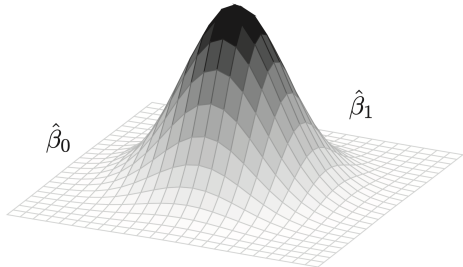


Figure 1 : Exemple illustratif d'une distribution a posteriori conjointe de l'ordonné à l'origine et de la pente. Une « marche aléatoire » dans l'espace des paramètres.

Fréquentiste vs Bayésien : deux philosophies

Approche fréquentiste :

- Les paramètres sont fixes mais inconnus
- Les données sont aléatoires
- Intervalles de confiance
- Tests d'hypothèses (valeurs p)

Interprétation difficile :

> *« Si l'hypothèse nulle était vraie, on observerait un effet au moins aussi extrême 5 % du temps »*

Approche bayésienne :

- Les paramètres sont incertains
- Les données sont fixes (observées)
- Intervalles de crédibilité
- Probabilités directes

Interprétation intuitive :

> *« Il y a 95 % de chances que le vrai paramètre soit dans cet intervalle »*

☞ Ouvrages de référence : (Kruschke *et al.* 2012 ; Kruschke 2015 ; Kruschke et Liddell 2018 ; McElreath 2020)

L'ANALYSE DE SENTIMENT : MADAME BOVARY

Tendances à travers les chapitres

Visualisation des patrons : une étape centrale dans notre analyse

☞ Sentiments à partir de l'extension `syuzhet` (dictionnaire NRC)* (Jockers 2015; Mohammad et Turney 2010)

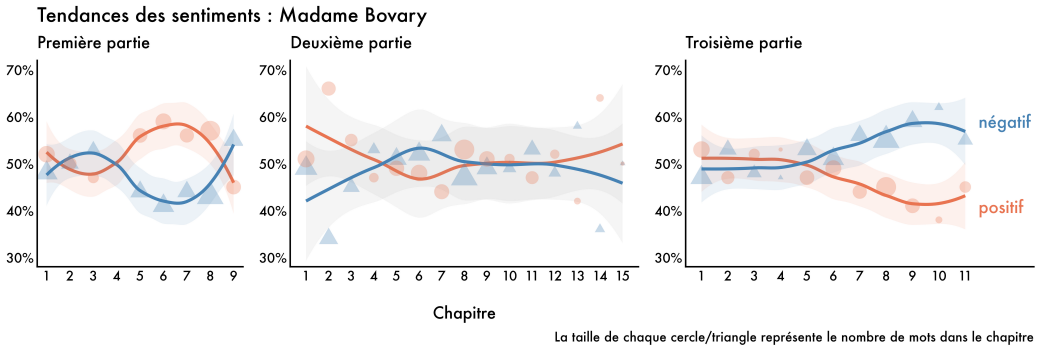


Figure 2 : Les courbes lissées ne reflètent pas les hypothèse du modèle à venir.

Tendances à travers les chapitres

Quel modèle choisir ?

- On a des options, mais aujourd'hui on explorera un **modèle binomial**
- C'est une **extension** d'une régression logistique : on modèle une **proportion**

☞ On compare deux modèles bayésiens :

Spécification des modèles : deux hypothèses dans les effets aléatoires

A. `nombre_positifs | trials(nombre_total) ~ 1 + (1 | chapitre)`

B. `nombre_positifs | trials(nombre_total) ~ 1 + (1 | partie / chapitre)`

☞ Les deux modèles supposent l'a priori $\beta_0 \sim \mathcal{N}(0, 1)$

Tendances à travers les chapitres

Deux perspectives

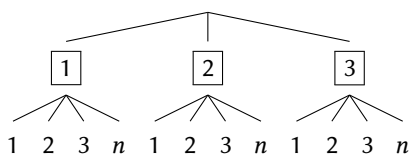
Les sentiments positifs dans les chapitres ...

A. varient de façon **indépendante** → parties = « étiquettes organisationnelles »

B. sont plus similaires dans la même partie qu'à travers les parties

A

B



Analyse 1

Modèles A et B

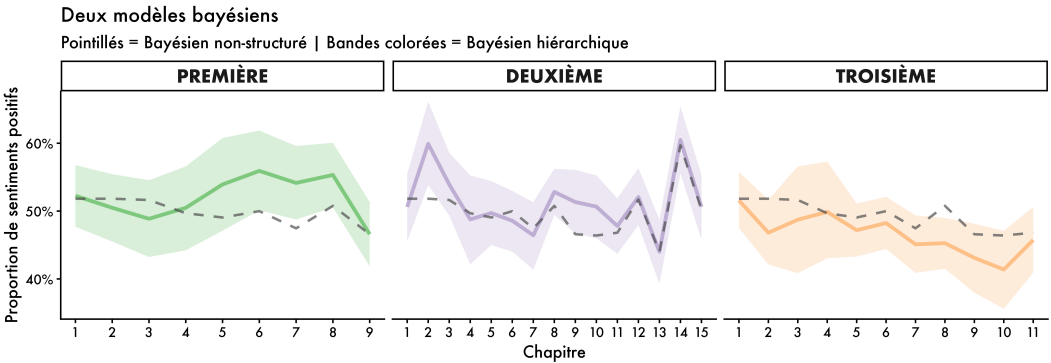


Figure 3 : Modèle non-structuré versus modèle hiérarchique (intervalles de crédibilité à 95 %).

Analyse 1

Comment évaluer les modèles

☞ LOO = *Leave-one-out cross validation*

Évalue la capacité prédictive du modèle en estimant sa performance sur des données non vues. Un score LOO plus élevé indique un **meilleur** précision prédictive

Modèle	ELPD diff	SE diff
Modèle hiérarchique	0.0	0.0
Modèle non-structuré	−24.8	9.0

TABLE 1 : Comparaison des modèles par validation croisée LOO

Analyse 1

Résultats et vérification prédictive

- Modèle **hiérarchique** : **meilleure** précision prédictive (≠ ELPD de 24,9 points)
- Structure hiérarchique (chapitres dans les parties) améliore la précision prédictive du modèle

☞ Chaque partie possède ses propres caractéristiques distinctives

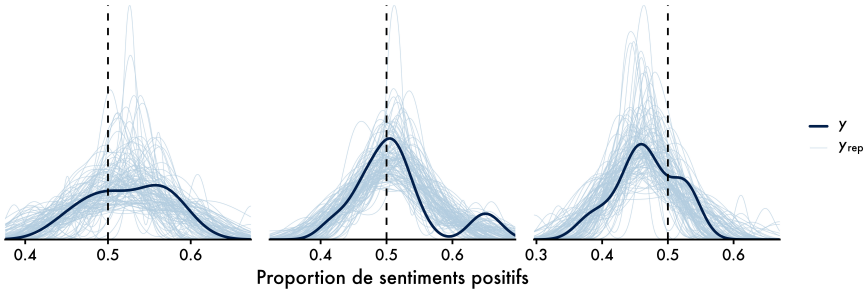


Figure 4 : Vérification prédictive du modèle (première, deuxième, troisième parties)

Analyse 1

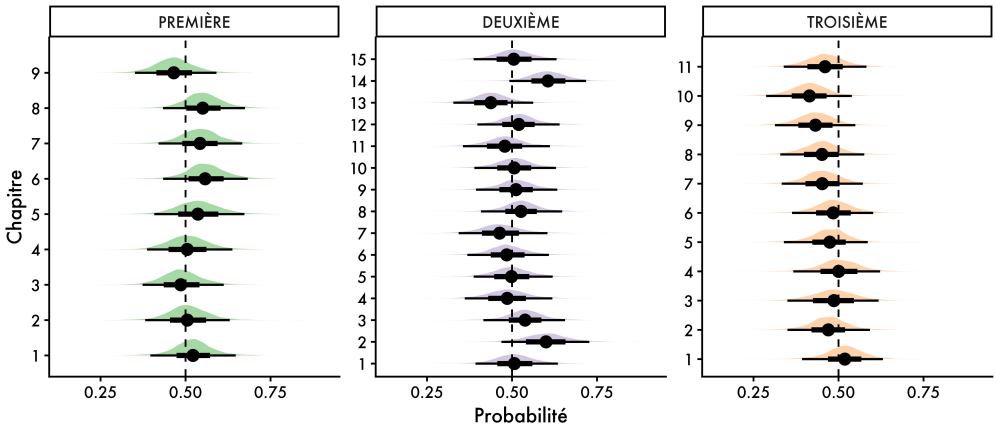


Figure 5 : Un résultat complet par rapport aux paramètres d'intérêt.

Analyse 1

L'analyse bayésienne nous donne :

1. Plus d'information sur les effets d'intérêt
2. Plus de **flexibilité** (distributions a priori, possibilité d'estimer la variance,[1] etc.)
3. Une interprétation **intuitive** (cf. intervalles de confiance)
4. Un degré d'incertitude beaucoup plus élevé (p. ex., vérifications prédictives)

☞ **Mais ses résultats sont-ils comparables à ceux d'une analyse fréquentiste ?**

[1] Utile pour estimer l'**homogénéité** des données (e.g., présentation de P. Schumacher aujourd'hui).

Analyse 2

Les parties ont-elles un effet ? Des comparaisons multiples

- Considérons deux modèles dont `partie` est une variable prédictive
  « Les parties peuvent-elles prédire la proportion des sentiments positifs ? »
- Un modèle **fréquentiste** et un modèle **bayésien**

Spécification du modèle

```
nombre_positifs | trials(nombre_total) ~ partie + (1 | chapitre)
```

Analyse 2

Comparaisons multiples

☞ Résultats similaires, mais pas identiques : intervalles de crédibilité ≠ intervalles de confiance

- La différence observée pourrait affecter catégoriquement nos conclusions

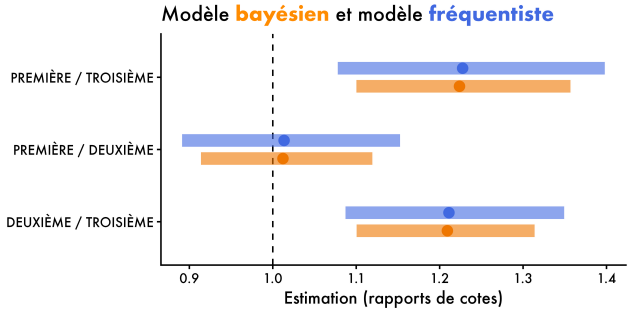


Figure 6 : Comparaisons multiples à partir des modèles.
Veuillez noter que les lignes oranges forment une **distribution** de probabilités.

Analyse 2

Comparaisons multiples

- Problème typique : erreurs de type 1 et taux d'erreur global → **corrections** (p. ex., Tukey)

☞ Dans une analyse bayésienne, cette notion n'existe plus (Voir Garcia 2025)

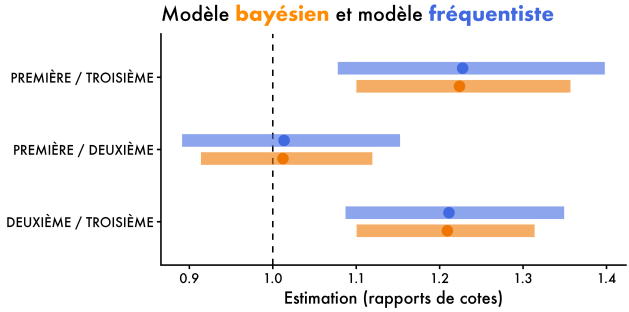


Figure 7 : Comparaisons multiples à partir des modèles.
Les comparaisons fréquentistes ont été ajustées selon la méthode Tukey.

Analyse 2

Comparaisons multiples

- Un modèle hiérarchique est déjà conçu pour réduire les erreurs de type I (Gelman et Tuerlinckx 2000)
- Les a priori exercent un effet de régularisation naturelle (diapo suivante)

☞ Cela stabilise les estimations et peut en améliorer la précision

- À noter que les ajustements traditionnels ne changent que les intervalles de confiance[2]

[2] Et les valeurs p en conséquence.

Analyse 2

Comparaisons multiples

- Notez que les estimations des modèles fréquentistes restent identiques :

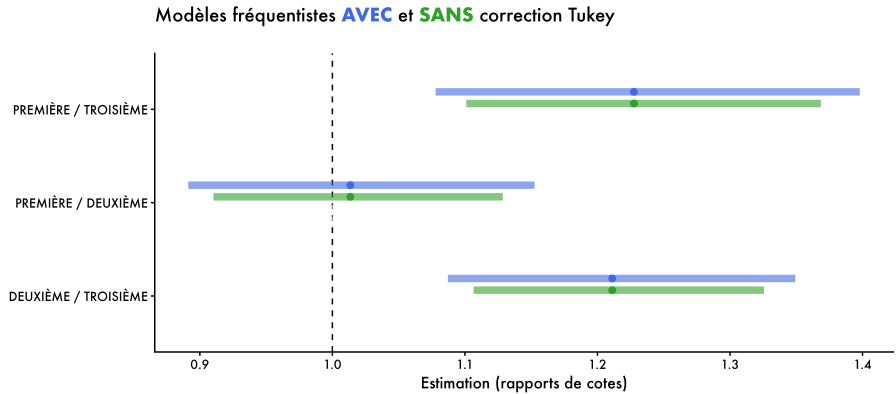


Figure 8 : Comparaisons additionnelles sans correction Tukey. Voir les lois a priori ici.

Analyse 2

Comparaisons multiples

- Les **a priori** fournissent un effet de régularisation supplémentaire :

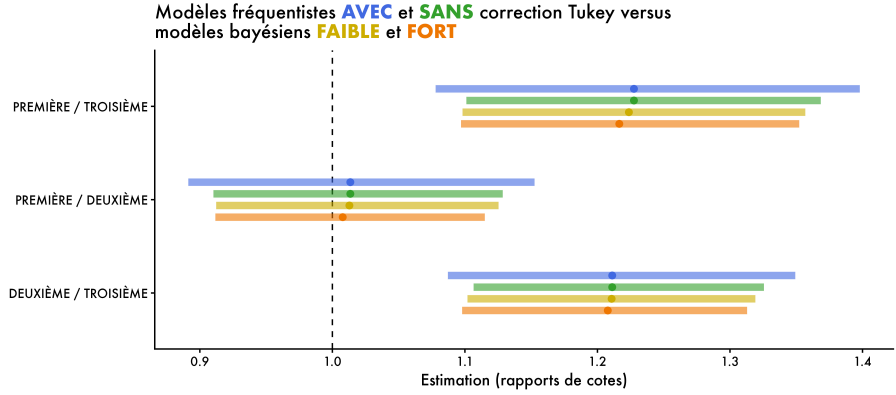


Figure 9 : Comparaisons additionnelles sans correction Tukey. Voir les lois a priori ici.

Analyse 2

- Au fait, un modèle bayésien nous donnes des **distributions** pour les paramètres

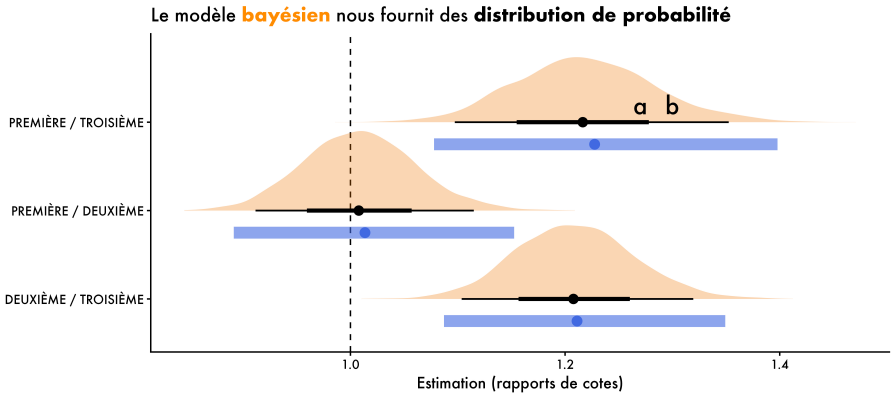


Figure 10 : Une différence importante : la valeur **a** est plus probable que la valeur **b**.

Conclusions

L'approche bayésienne : un atout pour les humanités numériques

- **Transparence** : Quantification explicite de l'incertitude vs. fausse précision de l'IA
- **Robustesse** : Les modèles hiérarchiques exploitent la structure des données littéraires
- **Intuitivité** : Énoncés probabilistes directs (« 95 % de chances que… »)
- **Flexibilité** : Pas de corrections *ad hoc* pour les comparaisons multiples

Points essentiels

- L'analyse bayésienne offre une voie rigoureuse et **transparente** pour analyser les textes tout en respectant l'incertitude inhérente à nos données.
- Une méthodologie qui permet de concilier sophistication statistique et honnêteté scientifique (p. ex., transparence, incertitude, connaissances a priori).

Références I

Guilherme D Garcia : Bayesian estimation in multiple comparisons. *Studies in Second Language Acquisition*, 2025. URL `https://gdgarcia.ca/multcomp`.

Andrew Gelman et Francis Tuerlinckx : Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational statistics*, 15(3) :373–390, 2000.

Matthew L. Jockers : *Syuzhet : Extract Sentiment and Plot Arcs from Text*, 2015. URL `https://github.com/mjockers/syuzhet`.

John K Kruschke : *Doing Bayesian data analysis : a tutorial with R, JAGS, and Stan*. Academic Press, London, 2nd édition, 2015.

John K Kruschke, Herman Aguinis et Harry Joo : The time has come : Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, 15(4) :722–752, 2012.

John K Kruschke et Torrin M Liddell : Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25 (1) :155–177, 2018.

Richard McElreath : *Statistical rethinking : a Bayesian course with examples in R and Stan*. Chapman & Hall/CRC, Boca Raton, 2nd édition, 2020.

Références II

Saif Mohammad et Peter Turney : Emotions evoked by common words and phrases : Using mechanical turk to create an emotion lexicon. *In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California, 2010. URL `http://saifmohammad.com/WebPages/lexicons.html`.

Merci !

☞ fr.gdgarcia.ca

guilherme.garcia@lli.ulaval.ca

A priori

Deux suppositions

☞ Dans quelle mesure sommes-nous certains d'un effet (c'est-à-dire quel est l'écart-type supposé dans les distributions normales a priori) ?

Modèle **faible** (moins certains) : Retourner

$\beta_0 \sim \mathcal{N}(0, 1)$

$\beta \sim \mathcal{N}(0, 0.5)$

Modèle **fort** (plus certains) :

$\beta_0 \sim \mathcal{N}(0, 0.25)$

$\beta \sim \mathcal{N}(0, 0.25)$

☞ En réalité, notre décision dépend de nos hypothèses et des connaissances déjà établies.