Analyse Statistique d'un Corpus Poétique

José Manuel Rodríguez Caballero

February 16, 2025

Contents

1 Description du Jeu de Données **1**
1.1 Étapes de constitution 2
1.2 Données manquantes 2

2 Présentation des Variables **2**
2.1 Niveau Cas-Témoin 2
2.2 Niveau Poète 2
2.3 Niveau Poème 3
2.4 Niveau Vers 3

3 Analyse avec Réduction de la Dimension **3**
3.1 Principe 3
3.2 Analyse en Composantes Principales (ACP) 3

1 Description du Jeu de Données

Le corpus étudié se compose de poèmes provenant de différents auteurs, dont certains sont reconnus comme suicidaires. Plusieurs étapes de nettoyage ont abouti à un ensemble final de données, organisé selon **quatre niveaux** :

1. **Cas-Témoin** : identifiés par `pair_id` (chaque poète suicidaire est apparié à un poète non suicidaire).
2. **Poète** : chaque auteur est décrit par diverses informations biographiques (dates, pays, orientation, etc.) et par un indicateur `suicidal`.
3. **Poème** : chaque recueil de vers possède une période (`Early`, `Middle`, `Later`), un titre, etc.
4. **Vers** : unité de base pour la mesure des émotions (colère, joie, tristesse, etc.).

1.1 Étapes de constitution

`raw_data.csv` Fichier de départ (42 lignes), chaque ligne représentant un poème, avec des métadonnées (dates, pays, lien source, etc.).

`clean_data_1.csv` Fichier intermédiaire où chaque vers est placé sur une ligne (2931 lignes au total). Les informations du poète sont dupliquées pour chaque vers du même auteur.

`clean_data_2.csv` Fichier final à granularité identique (1 vers par ligne), où le texte du vers est remplacé par des scores émotionnels (`anger`, `joy`, `sadness`, etc.).

1.2 Données manquantes

- Les 10 scores d'émotions ne comportent aucune valeur manquante.
- Au niveau Poète, la variable `heterosexual` contient 2 valeurs manquantes (`NA`).
- Les autres champs (dates, pays, etc.) sont complets.

Le nombre de valeurs manquantes étant très faible, on considère que cela n'entrave pas l'analyse.

2 Présentation des Variables

2.1 Niveau Cas-Témoin

- `pair_id` : identifiant de la paire (poète suicidaire vs. poète témoin).

2.2 Niveau Poète

- `poet` : nom de l'auteur (14 distincts).
- `suicidal` : `TRUE` ou `FALSE` (7 de chaque).
- `sex` : `Male` / `Female`.
- `heterosexual` : `TRUE` / `FALSE` / `NA`.
- `date_of_birth`, `date_of_death` : dates.
- `country_of_birth` : pays.

2.3 Niveau Poème

- `period` : `Early`, `Middle` ou `Later`.
- `poem_title` : titre du poème.

2.4 Niveau Vers

- `anger`, `anticipation`, `disgust`, `fear`, `joy`, `sadness`, `surprise`, `trust`, `negative`, `positive` : scores d'émotions.
- Chacune de ces variables est un compteur ou une pondération de mots associés à l'émotion concernée.
- Nombre d'observations : 2931

3 Analyse avec Réduction de la Dimension

3.1 Principe

L'analyse statistique se concentre sur les **moyennes d'émotions par poète**. Pour chaque auteur, on agrège (par la moyenne) les 10 scores d'émotions sur l'ensemble de ses vers, formant ainsi un vecteur dans $\mathbb{R}^{10}$. On obtient donc une matrice M de dimension 14×10 (14 poètes, 10 émotions).

3.2 Analyse en Composantes Principales (ACP)

Standardisation Avant l'ACP, on "centre-réduit" les 10 colonnes d'émotion, afin de les ramener à une moyenne nulle et un écart-type unitaire.

Décomposition Soient $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{10}$ les valeurs propres de la matrice de covariance de M. Les composantes principales (PC1, PC2, etc.) sont les vecteurs propres associés, ordonnés par importance décroissante.

Résultats

- **PC1 et PC2** : Les deux premières composantes expliquent environ 77,7 % de la variance (50,4 % pour PC1, 27,5 % pour PC2).
- **Position des poètes** : En projetant chaque poète dans le plan (PC1, PC2), on peut observer d'éventuelles tendances de regroupement selon `suicidal`. Aucune séparation claire n'est obligatoirement visible, mais des indices de différences émotionnelles peuvent se dégager.

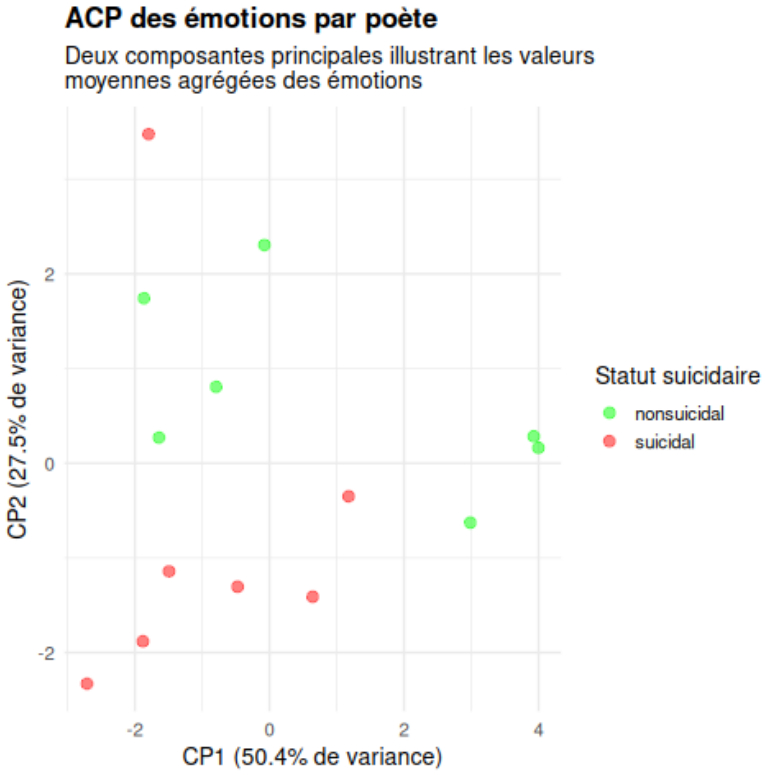

Figure 1: Représentation des poètes dans le plan défini par les deux premières composantes principales (PC1 et PC2). Les couleurs indiquent par exemple le statut suicidaire.