Analyse Statistique d'un Corpus Poétique

José Manuel Rodríguez Caballero

February 17, 2025

Contents

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 en Composantes Principales (ACP)

3.1 Principe et Préparation

L'objectif est de résumer les différences émotionnelles entre poètes à l'aide d'une **analyse en composantes principales (ACP)**. Pour chaque poète, on calcule la moyenne de ses scores d'émotion (`anger`, `joy`, `negative`, etc.), obtenant ainsi un vecteur dans $\mathbb{R}^{10}$. La matrice finale, de taille 14×10, est ensuite "centrée-réduite", de sorte que chaque variable ait une moyenne nulle et un écart-type unitaire.

3.2 Éboulis des Valeurs Propres (Scree Plot)

La Figure 1 présente l'éboulis des valeurs propres. Chaque barre correspond à la proportion de variance expliquée par une composante principale. Ici, on constate que les deux premières composantes expliquent déjà plus de 75 % de la variance cumulée. Cela motive l'examen plus approfondi des axes PC1 et PC2.

3.3 Projection des Poètes (PC1, PC2)

Dans la Figure 2, on projette les poètes (individus) dans le plan formé par les deux premières composantes principales. Les points sont colorés en fonction du statut `suicidal` (`TRUE` ou `FALSE`).

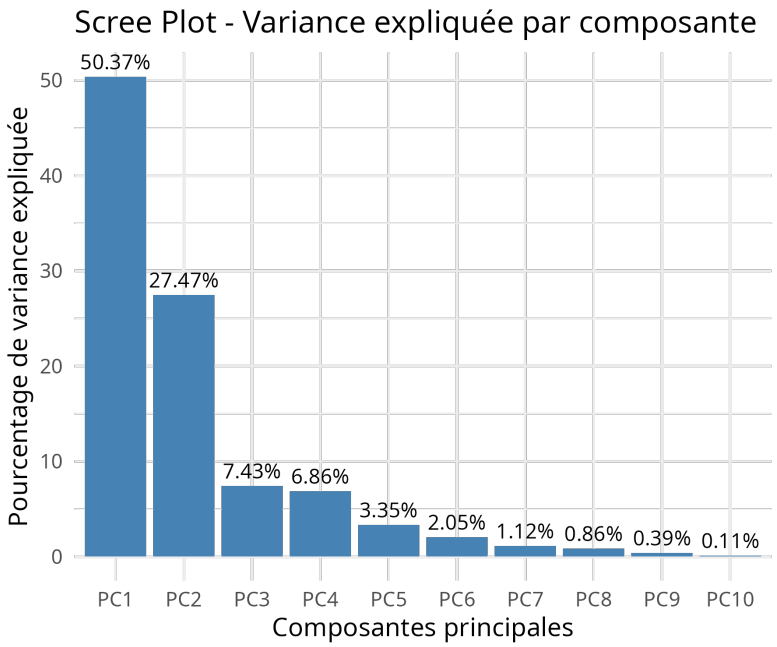

Figure 1: Éboulis des valeurs propres : part de la variance expliquée par chaque composante (PC1 à PC10).

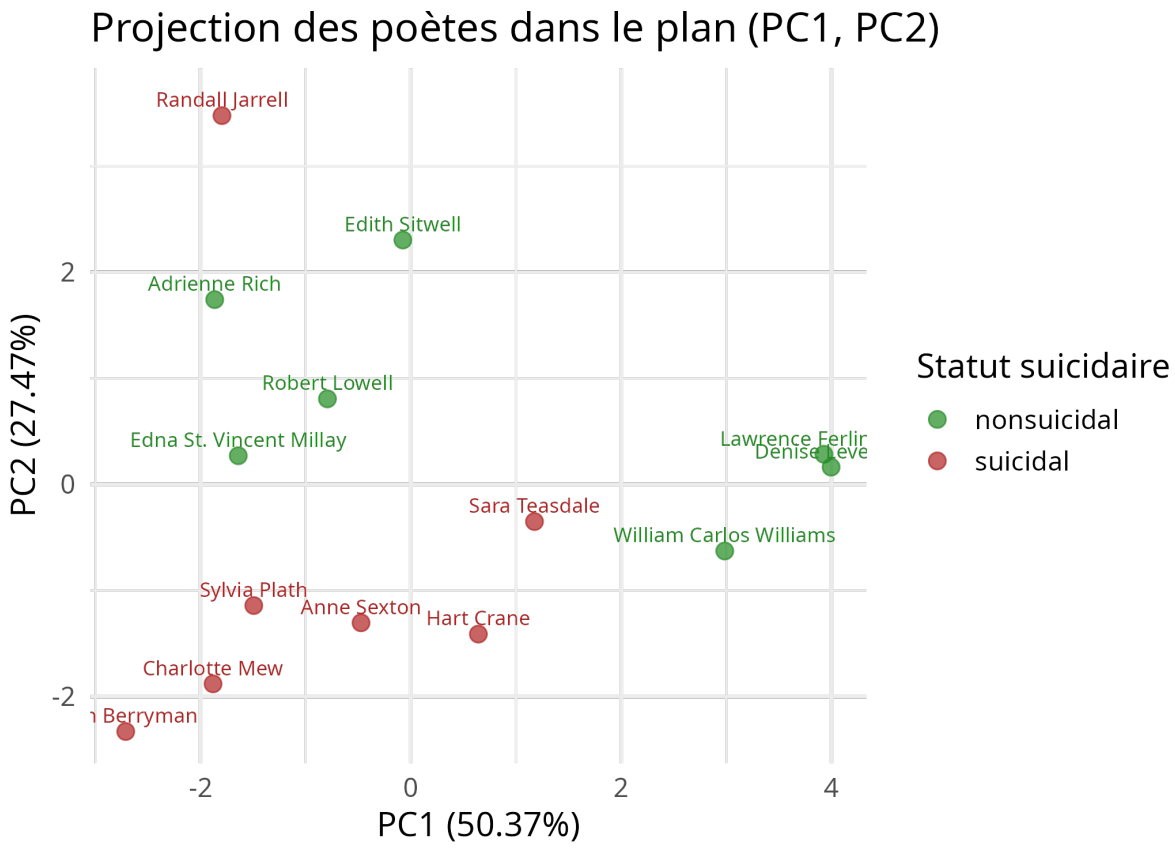

Figure 2: Projection des poètes sur les deux premières composantes principales (PC1 et PC2).

Même si l'on observe une dispersion relativement importante, on peut investiguer si les poètes suicidaires se situent dans une zone distincte. Nous discutons plus loin (Section 3.5) de la comparaison statistique des scores de PC1 et PC2 selon le statut suicidaire.

3.4 Visualisation des Variables (Biplot et Cercle de Corrélations)

Biplot. La Figure 3 superpose la projection des individus (poètes) et des variables (émotions) dans le même plan. Les flèches indiquent la direction d'accroissement d'une émotion particulière, et leur longueur traduit l'importance de la représentation sur PC1–PC2. Un poète situé dans la même direction qu'une flèche présente une valeur élevée pour cette émotion (par rapport aux autres).

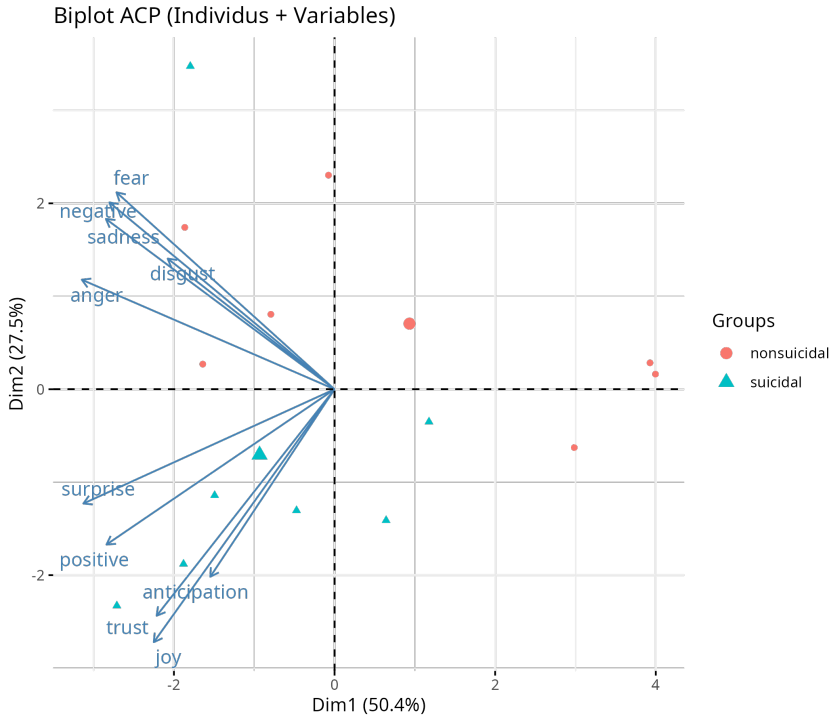

Figure 3: Biplot ACP : individus et variables (émotions) dans le plan (PC1, PC2).

Cercle des Corrélations. La Figure 4 illustre uniquement les émotions (variables) et leurs corrélations dans l'espace des deux premières composantes. Deux variables dont les vecteurs forment un angle faible sont positivement corrélées (et inversement pour un angle proche de 180°). Plus le vecteur est long, plus la variable est bien représentée par PC1–PC2.

3.5 Comparaison Statistique des Scores sur PC1 et PC2

Pour évaluer si la position des poètes sur les deux premiers axes diffère significativement selon le statut `suicidal`, on compare ci-après les distributions de PC1 et PC2 par groupe.

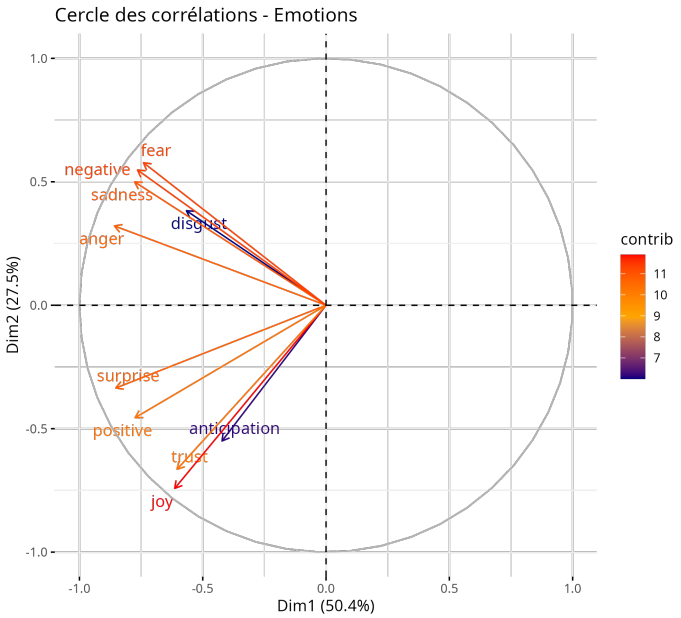

Figure 4: Cercle des corrélations : variables émotionnelles projetées sur PC1 et PC2.

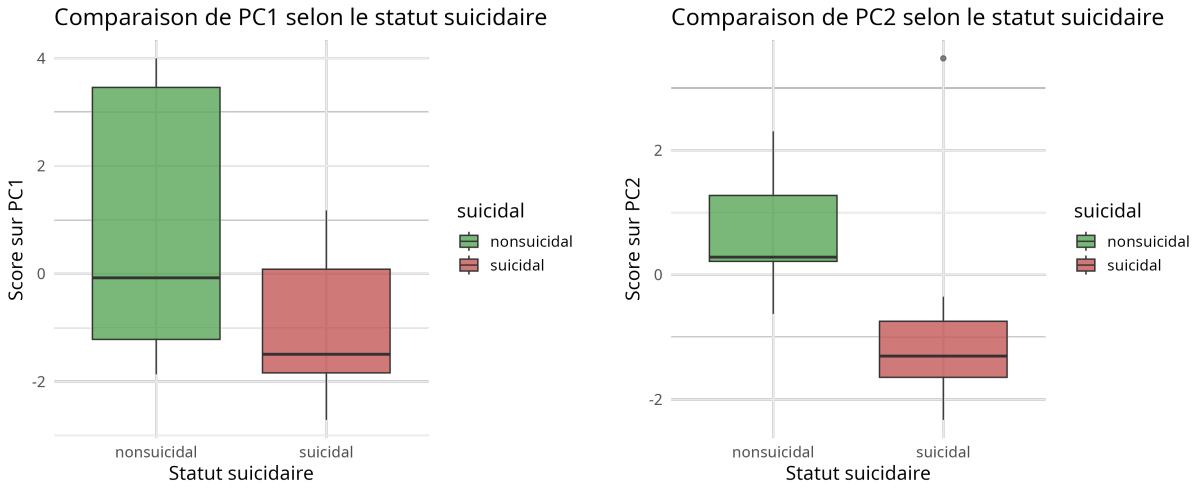

Figure 5: Boxplots comparant PC1 (gauche) et PC2 (droite) selon le statut suicidaire.

Les boxplots (Figure 5) montrent qu'aucun groupe ne se détache de façon flagrante. Afin d'étayer cela, on effectue un *Welch Two Sample t-test* pour chacune des composantes (PC1 et PC2). Les résultats se trouvent dans le fichier `07_tests_statistiques.txt` et sont reproduits ci-dessous.

Test sur PC1.

```
t = 1.6581, df = 9.2921, p-value = 0.1306
IC 95%: [-0.6681746, 4.4029684]
Moyenne (nonsuicidal) = 0.9336984
Moyenne (suicidal)    = -0.9336984
```

Test sur PC2.

```
t = 1.7063, df = 9.0007, p-value = 0.1221
IC 95%: [-0.4597469, 3.2824927]
Moyenne (nonsuicidal) = 0.7056865
Moyenne (suicidal)    = -0.7056865
```

Dans les deux cas, la *p-value* est supérieure à 0,05, indiquant qu'il n'y a pas de différence statistiquement significative entre poètes suicidaires et non suicidaires concernant les scores moyens sur PC1 ou PC2. On note cependant que les moyennes présentent des signes inverses ("positifs" pour les poètes non suicidaires et "négatifs" pour les suicidaires). Ce "renversement" n'est pas suffisant pour être jugé significatif au vu de l'échantillon restreint ($p\text{-}value > 0{,}1$).

En conclusion, l'ACP permet de visualiser les différences (ou similarités) dans l'espace des émotions, mais elle ne révèle pas, selon ces tests, de séparation nette entre poètes suicidaires et non suicidaires sur les deux premières composantes.