

INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

M2 ACTUARIAT
ANNÉES 2023/2024

ESTIMATION DE COPULES

Projet 2 : Étude de la structure de dépendance, Spotify data

Léo ALLERS
Théo JALABERT
Lou SIMONEAU-FRIGGI

Table des matières

Introduction	2
I Présentation des données	3
I.1 Jeu de données	3
I.2 Description qualitative des données	3
I.3 Outils de détection de la copule et de la structure de dépendance	4
II Estimation paramétrique	9
II.1 Estimation des marginales	9
II.2 Estimation de la copule et des paramètres	10
II.3 Qualité de l'ajustement paramétrique	13
III Estimation non-paramétrique	15
III.1 Estimation des paramètres de la copule	15
III.2 Qualité de l'ajustement non-paramétrique	18
Conclusion	20
Bibliographie	21
Annexes	22

Introduction

Dans un monde où la musique joue un rôle central dans notre quotidien, comprendre les intrications subtiles entre les caractéristiques des chansons est devenu essentiel pour les artistes, les producteurs, ainsi que les plateformes de streaming telle que Spotify. Ce rapport se concentre sur l'analyse de la structure de dépendance entre deux attributs fondamentaux des œuvres musicales : « **Danceability** » et la « **Speechiness** ». La « Danceability » reflète le degré auquel un titre est propice à la danse, basé sur des éléments comme le tempo, la régularité rythmique et l'intensité. D'autre part, la « Speechiness » mesure la présence de paroles dans une chanson, offrant un aperçu unique sur le ratio entre musique et parole.

L'étude de la relation entre ces deux variables est non seulement fascinante du point de vue de l'analyse musicale, mais elle revêt également une importance commerciale et culturelle significative. Elle permet de mieux comprendre comment les caractéristiques d'une chanson influencent sa réception par le public, et peut guider les artistes dans la création de morceaux qui répondent aux préférences des auditeurs. De plus, cette analyse offre des perspectives précieuses aux plateformes de streaming pour affiner leurs algorithmes de recommandation, améliorant ainsi l'expérience utilisateur. De telles analyses peuvent également être intéressantes pour des entreprises comme **One28**, fondée par un Isfaïen et ami, Adrien CRASTES, dans la mesure où nous pourrions étudier la réceptivité du public à certains styles de musique, permettant ainsi aux DJ et aux établissements de nuit d'adopter des stratégies afin de fidéliser et de cibler leur clientèle.

En utilisant la théorie des copules pour explorer la structure de dépendance entre la "Danceability" et la "Speechiness", ce rapport vise à dévoiler des liens cachés dans les données de Spotify, ouvrant ainsi la voie à une compréhension plus profonde de la musique que nous aimons.

I Présentation des données

Dans le cadre de ce projet, nous allons mettre en oeuvre plusieurs méthodes afin de déterminer la copule la plus adaptée à la description de la structure de dépendances entre nos deux variables quantitatives.

I.1 Jeu de données

Le jeu de données utilisé pour ce projet est accessible sur le site Kaggle¹, où l'on peut trouver six ensembles de données distincts. Chacun de ces ensembles contient des informations détaillées sur des titres musicaux diffusés sur la plateforme Spotify, classés par décennie. Pour notre analyse, nous avons choisi de nous concentrer sur les titres de la décennie 2010, en mettant particulièrement l'accent sur l'étude de la relation de dépendance entre deux variables spécifiques :

- *Danceability* : Cette variable évalue la facilité avec laquelle on peut danser sur un titre musical. Elle est calculée à partir de divers critères, tels que le tempo, la régularité du rythme et l'intensité de la musique. Les valeurs de cette variable varient de 0 à 1, où 0 indique un titre non adapté à la danse, tandis que 1 représente une chanson parfaitement adaptée à la danse.
- *Speechiness* : Cette variable mesure la quantité de paroles présentes dans une chanson. Elle est évaluée sur une échelle allant de 0 à 1, où une valeur proche de 1 indiquerait un titre avec un contenu vocal prédominant, tel qu'un poème récité, tandis qu'une valeur proche de 0 suggérerait une absence de paroles, comme c'est le cas pour un morceau instrumental classique.

L'objectif principal de ce projet est donc de déterminer une copule qui modélise de manière adéquate la corrélation entre l'envie de danser sur un titre musical des années 2010 et la présence de paroles dans celui-ci. En identifiant la copule la plus adaptée, nous pourrons mieux comprendre la nature de la dépendance entre ces deux aspects importants de la musique contemporaine.

I.2 Description qualitative des données

Dans cette analyse préliminaire, nous observons des caractéristiques distinctes pour nos deux variables principales, *Danceability* (capacité à danser) et *Speechiness* (quantité de texte), dans les musiques des années 2010.

• ***Danceability*** : La distribution de la variable *Danceability* montre que la majorité des titres musicaux de cette décennie offrent une bonne capacité à danser. En effet, une grande proportion des données se situe entre 45% et 70%, indiquant que la plupart des chansons sont relativement adaptées à la danse. La moyenne de *Danceability* se situe autour de 56%, ce qui renforce cette tendance vers des titres favorables à la danse.

• ***Speechiness*** : À l'opposé, la variable *Speechiness* présente une distribution nettement différente. La majorité des titres musicaux contiennent relativement peu de paroles,

1. Kaggle - Spotify Database link

avec une grande partie des données se situant en dessous de 15%. La moyenne de *Speechiness* est seulement de 9%, suggérant que la plupart des chansons de cette époque sont principalement musicales avec peu de contenu verbal. Cependant, il est important de noter la présence de certaines chansons ayant une proportion de paroles significativement plus élevée, allant jusqu'à un maximum de 96%, ce qui indique une variété dans le type de contenu vocal des titres.

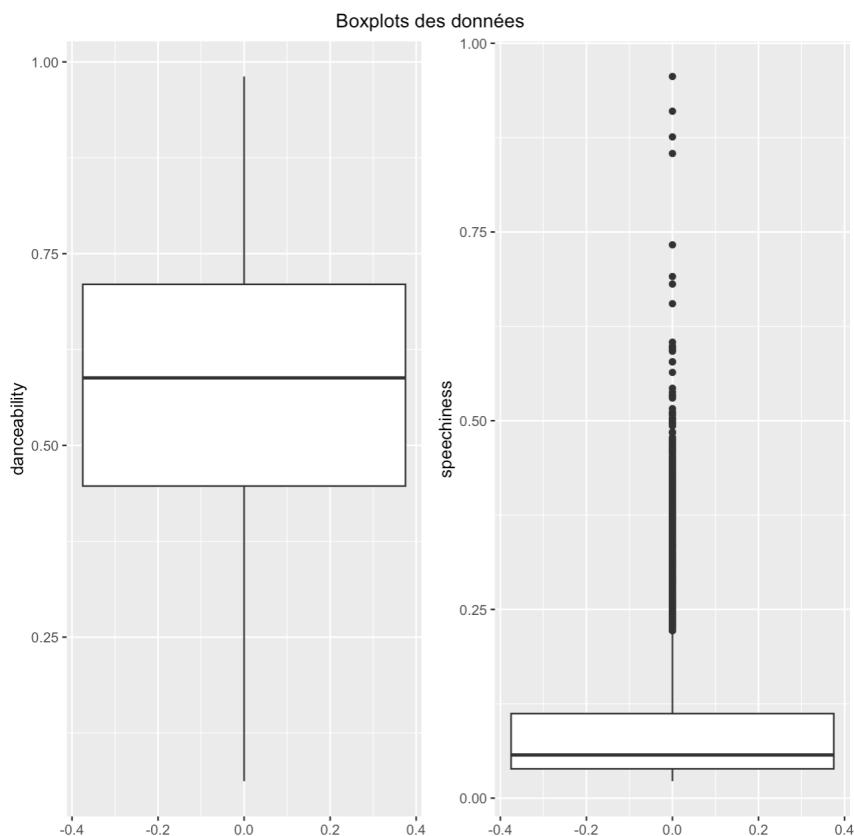


FIGURE 1 – Boîte à moustache des variables d'intérêt

Cette analyse descriptive met en évidence des différences marquées dans la répartition de ces deux caractéristiques au sein des musiques des années 2010, posant ainsi les bases pour une étude plus approfondie de la structure de dépendance entre ces deux variables.

I.3 Outils de détection de la copule et de la structure de dépendance

Le premier outil que nous utilisons pour analyser la relation entre *Danceability* et *Speechiness* est le diagramme de dispersion. Ce graphique fournit une représentation visuelle des données, permettant d'observer les patterns de corrélation entre les deux variables. Cependant, il est important de souligner que le diagramme de dispersion présente certaines limites dans la détection de la structure de dépendance spécifique entre les variables.

Diagramme de dispersion associé à l'échantillon

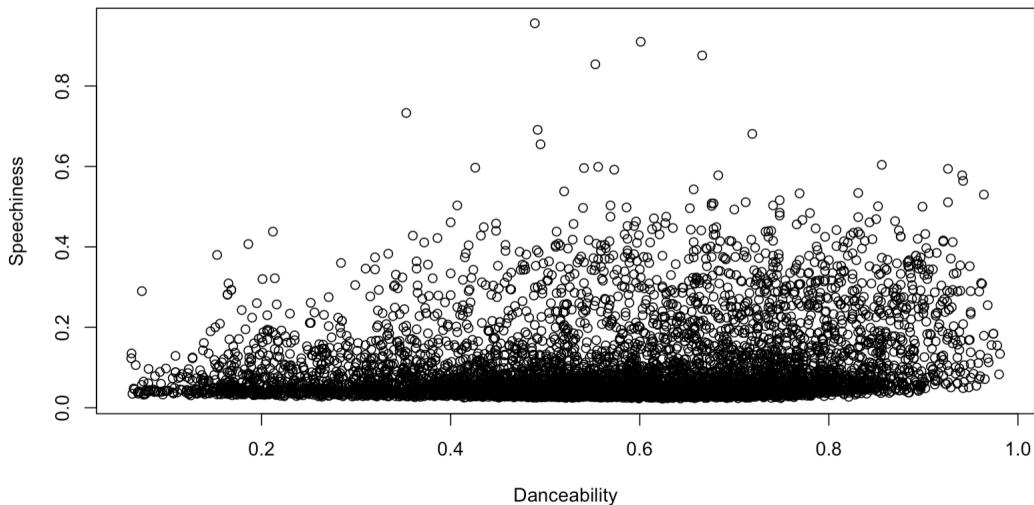


FIGURE 2 – Diagramme de dispersion

La FIGURE 2 présente un surplus d'informations proche de l'axe des abscisses.

L'une des principales limitations du diagramme de dispersion est qu'il inclut des informations sur les distributions marginales de chacune des variables d'intérêt, *Danceability* et *Speechiness*, en plus de leur relation de dépendance. Cela signifie que le diagramme reflète non seulement la façon dont les variables interagissent entre elles, mais également leurs caractéristiques individuelles. Par conséquent, il peut être difficile de discerner la nature exacte de la dépendance entre les variables à partir de ce graphique seul.

Pour surmonter cette limite, il est essentiel d'employer des outils statistiques plus sophistiqués qui se concentrent spécifiquement sur la dépendance. Ces outils peuvent inclure l'utilisation de copules pour modéliser la dépendance conjointe des variables, indépendamment de leurs marginales. En appliquant des méthodes telles que l'ajustement de copules, les tests de bonne adéquation (*goodness-of-fit*) pour différentes familles de copules, et l'analyse de la dépendance conditionnelle, nous pouvons obtenir une compréhension plus précise de la structure de dépendance entre *Danceability* et *Speechiness*.

Nous pouvons tracer un histogramme en 3D, présenté par la FIGURE 3 ci-dessous, afin d'avoir une idée de la densité de la copule que l'on cherche à déterminer.

Histogramme 3D associé à l'échantillon

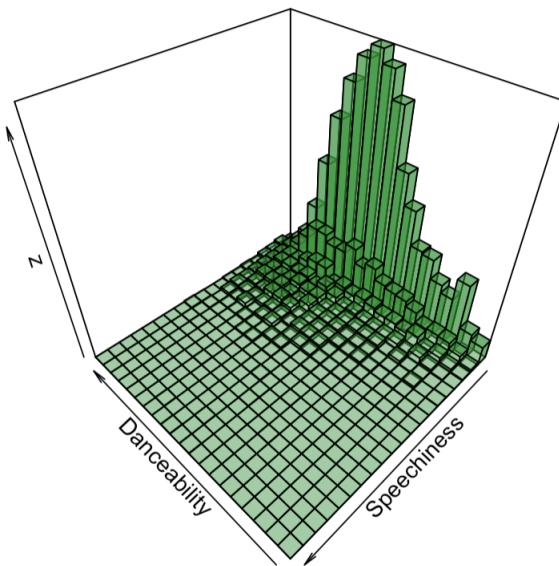


FIGURE 3 – Histogramme 3D sur nos données

Ici, il n'est pas évident d'identifier une copule classique qui pourrait être adaptée à nos données. Dès lors, il peut être pertinent de tracer un Rank-Rank plot, qui devrait être davantage adapté à notre étude.

Rank-Rank plot associé à l'échantillon

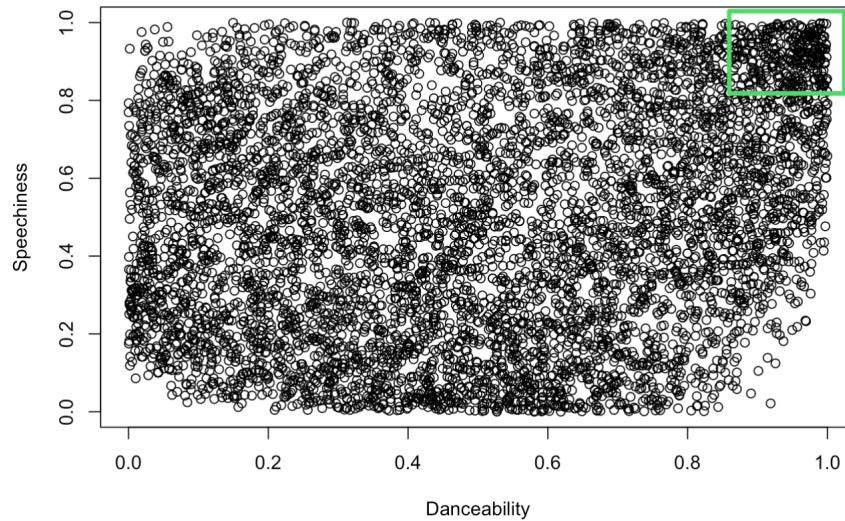


FIGURE 4 – Rank-Rank Plot

Nous pouvons observer une dépendance forte des extrêmes à droite. De plus, peu de valeurs se situent en bas à droite et en bas à gauche. Cela nous fait donc penser à une copule de Gumbel qui présente, a priori, des caractéristiques similaires à notre échantillon.

Par ailleurs, il est pertinent de tracer la copule empirique de nos données (voir FIGURE 5 ci-après).

Copule empirique associée à l'échantillon

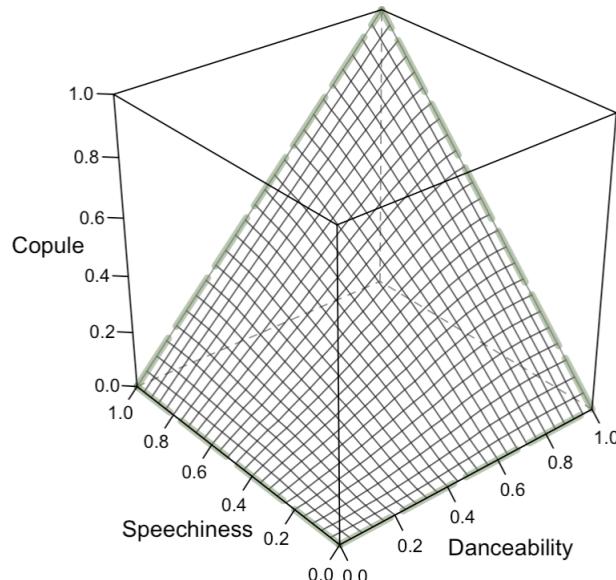


FIGURE 5 – Copule empirique

Cette dernière met en évidence un pic pour les valeurs élevées des variables, en revanche, elle manque de précision et ne permet pas de distinguer d'autres zones de forte dépendance. Dès lors, après avoir étudié les graphiques précédents, il est primordial d'utiliser quelques outils supplémentaires, non graphiques, pour détecter les dépendances. Le coefficient linéaire de Pearson et les coefficients non paramétriques de corrélation de Kendall et de Spearman sont présentés dans le tableau ci-dessous.

Pearson	τ de Kendall	ρ de Spearman
0,2000905	0,1278142	0,1953521

FIGURE 6 – Coefficients de corrélation

Ces trois coefficients démontrent une dépendance positive entre nos deux variables. De plus, ils sont plus proches de 0 que de 1, impliquant donc une corrélation relativement faible.

Le K-plot suivant vient appuyer ces deux remarques. En effet, l'ensemble des points se situent au dessus de la diagonale et en sont relativement proches.

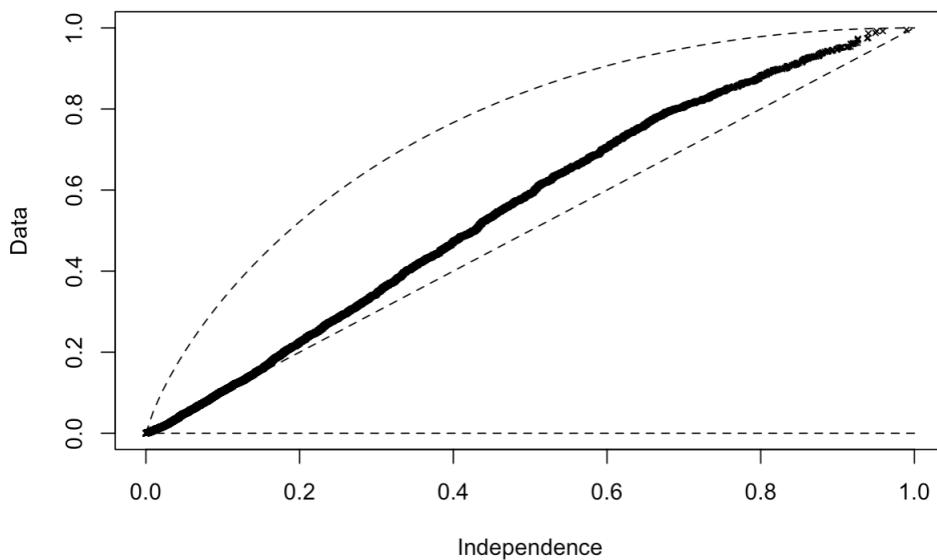


FIGURE 7 – K-plot associé à l'échantillon

De même, le Khi-plot ci-dessous, FIGURE 8, met en avant une dépendance entre nos deux variables d'intérêt puisque l'ensemble des points ne se situent pas à l'intérieur de l'intervalle tracé au voisinage de 0. Toutefois, les coefficients χ restent relativement proche de 0, ce qui traduit une faible dépendance.

Khi-plot associé à l'échantillon

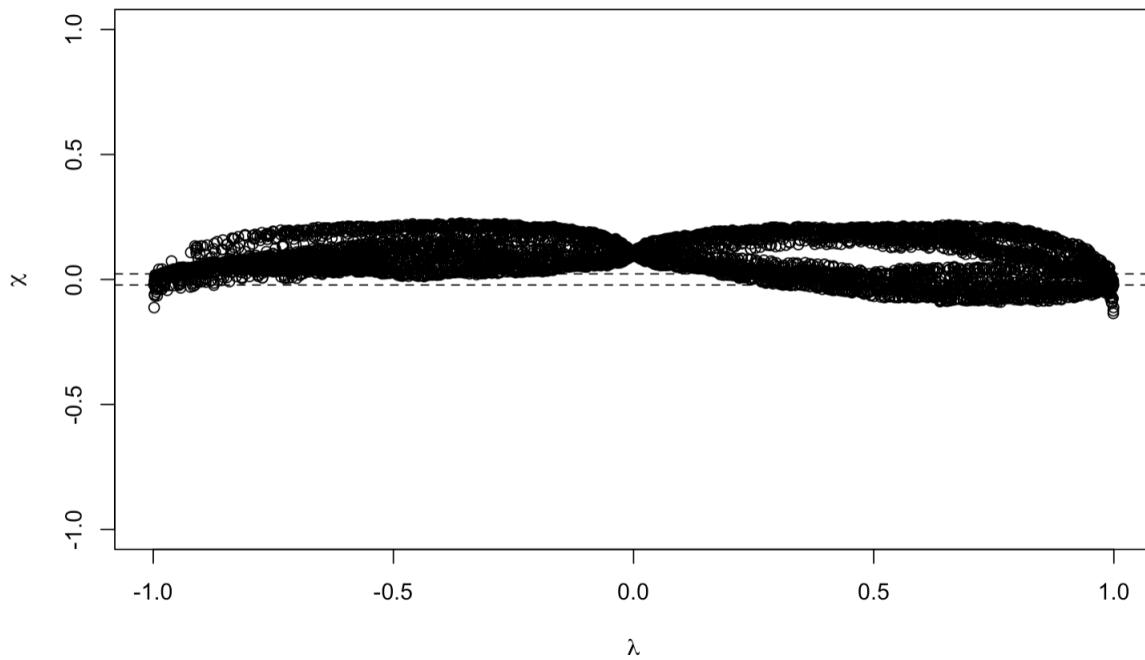


FIGURE 8 – Khi-plot associé à l'échantillon

Maintenant que nous avons réalisé ces premières analyses, nous allons estimé la copule associée à nos données.

II Estimation paramétrique

Dans un premier temps, nous allons nous intéresser aux différentes méthodes paramétriques d'estimation de copule.

II.1 Estimation des marginales

Dans cette phase de l'étude, une hypothèse a été formulée concernant les distributions marginales des variables analysées. Une fonction de R a été employée pour sélectionner automatiquement les distributions les plus adaptées selon le critère d'Information d'Akaike (AIC). Cette sélection a été effectuée parmi dix-huit distributions différentes, incluant des lois telles que Gumbel, Exponentielle, Gamma inverse, Gaussienne inverse, Pareto et Weibull.

En fin de compte, la loi Kumaraswamy et la loi Weibull inverse ont été retenues pour ajuster respectivement la variable *Danceability* et *Speechiness*. Les paramètres de ces lois ont été estimés par la méthode du maximum de vraisemblance. Ces paramètres, ainsi que les détails de l'ajustement, sont présentés dans le tableau ci-après.

<i>Danceability</i>		<i>Speechiness</i>	
Loi Weibull inverse		Loi Kumaraswamy	
a	b	Shape	Rate
2,766	2,794	1,794	19,966

FIGURE 9 – Paramètres des lois marginales

Les histogrammes en fréquence présentés ci-dessous jouent un rôle clé dans la validation de notre sélection des distributions marginales. En superposant les histogrammes des distributions ajustées (représentés en noir hachuré) aux histogrammes des données de l'échantillon (représentés en vert), nous pouvons observer visuellement à quel point ces distributions coïncident.

L'alignement étroit entre les histogrammes ajustés et ceux de l'échantillon suggère une adéquation significative entre les distributions marginales choisies et les caractéristiques réelles des données. Cette cohérence est un indicateur fort que les lois Kumaraswamy et Weibull inverse sont bien appropriées pour modéliser respectivement les variables que nous avons étudiées.

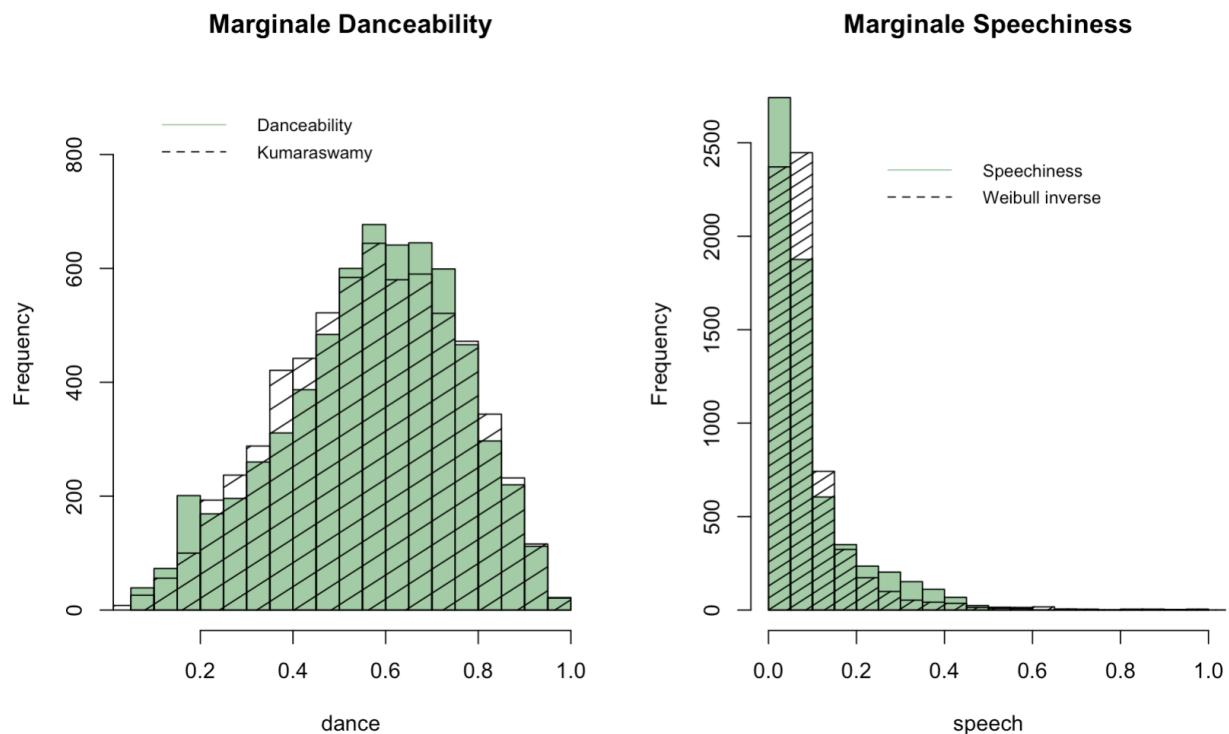


FIGURE 10 – Estimation des marginales

Dès lors, à partir de cette hypothèse sur les marginales, nous pouvons estimer la copule ainsi que ses paramètres.

II.2 Estimation de la copule et des paramètres

Pour déterminer la copule la plus adaptée à nos données, il est crucial d'abord d'examiner une variété de types de copules. Cette sélection initiale permet de couvrir un large éventail de structures de dépendance potentielles. Chaque copule candidate est ensuite ajustée aux données à l'aide d'une fonction telle que *fitCopula* en R.

Une fois les copules ajustées, le processus de sélection se concentre sur deux critères clés : le critère d'Information d'Akaike (AIC) et la log-vraisemblance. L'AIC est utilisé pour évaluer la qualité de l'ajustement de chaque copule en tenant compte du nombre de paramètres du modèle, avec une préférence pour les modèles qui minimisent l'AIC. La log-vraisemblance, quant à elle, mesure la probabilité que la copule ajustée produise les données observées, avec une préférence pour les modèles qui maximisent cette valeur.

Les copules finalement sélectionnées sont celles qui atteignent un équilibre optimal entre un faible AIC et une log-vraisemblance élevée, indiquant ainsi un ajustement efficace et parcimonieux aux données. Les résultats obtenus pour ces deux critères, pour chaque copule candidate, sont présentés dans le tableau ci-dessous. Ce tableau résume les performances de chaque copule, facilitant ainsi la comparaison et la sélection finale de la copule la plus appropriée pour modéliser la dépendance entre les variables étudiées.

Copule	Gaussienne	T	Joe	Gumbel	Franck	AMH
AIC	-284,6072	-281,0594	-376,3379	-341,6101	-226,9256	-177,3035
Log vraisemblance	143,3	142,5	189,2	171,8	114,5	89,65

FIGURE 11 – Critères de sélection des copules ajustées à notre échantillon

D'après les analyses et les critères de sélection que nous avons utilisés, la copule de Joe apparaît comme étant la plus adaptée à nos données, suivie de près par la copule de Gumbel. Cette conclusion est en accord avec nos observations préliminaires basées sur le rank-rank plot, où nous avions envisagé que la copule de Gumbel pourrait être une candidate appropriée.

Le choix de la copule de Joe comme la plus adaptée est particulièrement intéressant. Cette copule est reconnue pour sa capacité à modéliser efficacement les dépendances de queue supérieure, ce qui peut être pertinent dans le contexte de nos variables d'intérêt. La copule de Gumbel, quant à elle, est bien adaptée pour capturer les dépendances asymétriques, en particulier dans la queue supérieure.

Les paramètres de ces deux copules ajustées, qui ont été estimés par la méthode du maximum de pseudo-vraisemblance, sont résumés dans le tableau ci-dessous. L'utilisation du maximum de pseudo-vraisemblance est appropriée ici, car elle fournit une méthode d'estimation efficace en présence de données multivariées.

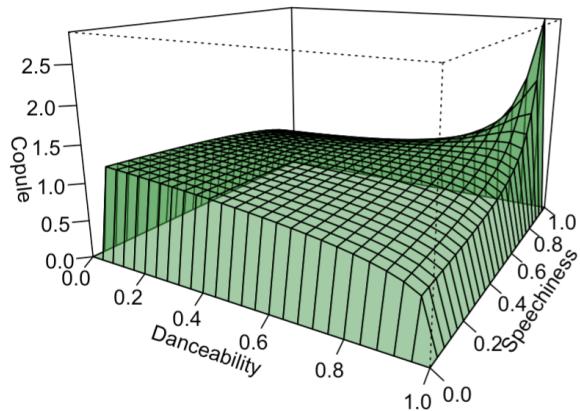
Ces paramètres estimés sont essentiels pour la suite de notre analyse. Ils vont nous permettre de tracer les densités des copules sélectionnées et d'évaluer la qualité de leur ajustement par rapport aux données. Cette étape est cruciale pour confirmer que les copules choisies reflètent fidèlement la structure de dépendance sous-jacente dans nos données.

Copule	Joe	Gumbel
Paramètre	1,219	1,139

FIGURE 12 – Paramètres des copules

Les deux copules ci-dessus ont des propriétés similaires comme une forte dépendance des extrêmes à droite et présentent des densités à peu près semblables. La seul point de discordance, semble être au niveau de la dépendance des extrêmes à gauche, comme on peut le voir avec la FIGURE 13.

Densité de la Joe ajustée



Densité de la Gumbel ajustée

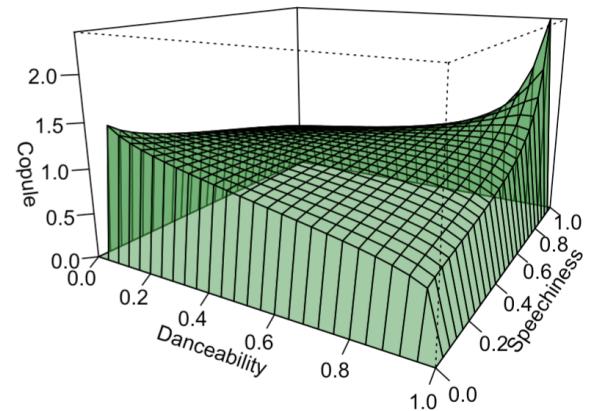


FIGURE 13 – Comparaison des résultats paramétriques

Cette analyse paramétrique indique que la copule de Joe est la mieux adaptée pour représenter la dépendance entre nos deux variables. Cette conclusion est renforcée par l'observation de la dépendance des queues de distribution, aussi connue sous le nom de "tail dependence".

La tail dependence est un concept clé en théorie des copules, qui se concentre sur la dépendance existante entre les valeurs extrêmes des variables. En d'autres termes, elle évalue dans quelle mesure des valeurs extrêmement élevées (ou faibles) d'une variable sont susceptibles de coïncider avec des valeurs extrêmement élevées (ou faibles) de l'autre variable. La copule de Joe est particulièrement adaptée pour modéliser les dépendances de queue supérieure, ce qui signifie qu'elle peut efficacement capturer la relation entre les événements extrêmes dans les données.

Le graphique de la tail dependence pour nos variables d'intérêt apporte un support visuel à cette conclusion. En comparant la dépendance des queues dans nos données avec celle théoriquement induite par une copule de Joe, on observe une correspondance étroite. Cela suggère que la copule de Joe capture fidèlement non seulement la dépendance globale entre les variables, mais aussi et surtout leur comportement conjoint dans les extrêmes.

Cette concordance entre les données empiriques et les caractéristiques de la copule de Joe confirme l'adéquation de ce choix de copule pour modéliser la structure de dépendance entre nos variables. Ainsi, la copule de Joe s'avère être un outil précieux pour comprendre et représenter la relation complexe entre les aspects de *danceability* et de *speechiness* dans les données de musique.

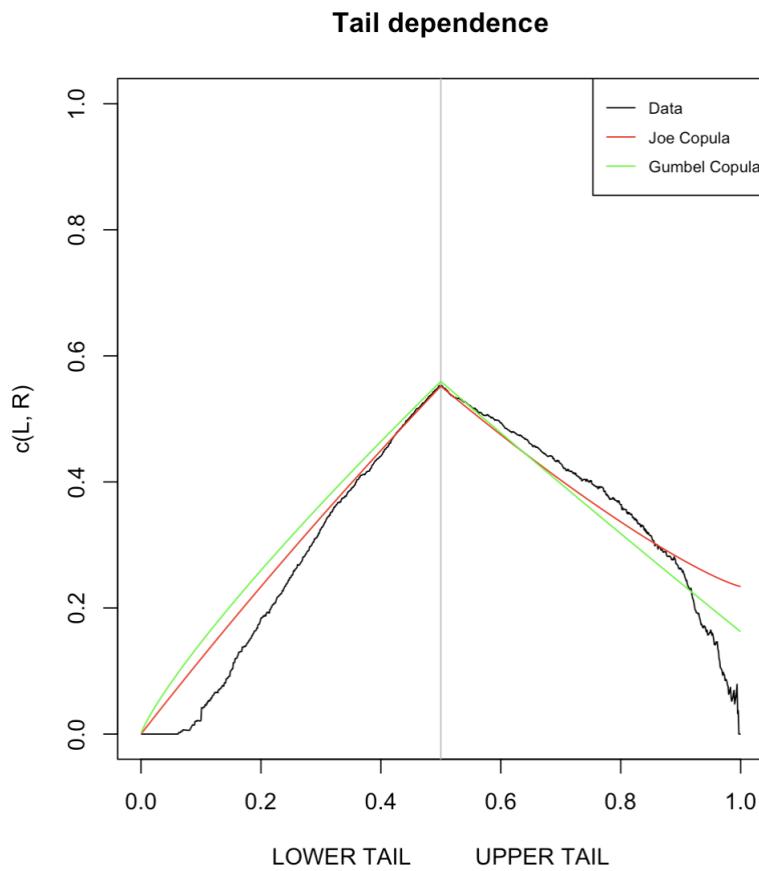


FIGURE 14 – Tail dependence

Afin de confirmer plus solidement le choix d'une copule de Joe, il est nécessaire d'avoir recours à des tests statistiques en complément.

II.3 Qualité de l'ajustement paramétrique

Dans notre étude, nous avons opté pour l'utilisation du test d'ajustement de Cramer-Von-Mises, implémenté dans la fonction **biCopulaGOF**, pour évaluer l'adéquation des copules à nos données. Ce choix s'inscrit dans le cadre de notre analyse visant à déterminer la copule la plus pertinente pour modéliser la dépendance entre les variables d'intérêt.

Les résultats de ce test montrent que les p-values pour les copules de Joe et de Gumbel sont toutes deux inférieures à 5%. Cette observation indique que les hypothèses de bon ajustement pour ces copules ne sont pas rejetées, suggérant ainsi que les deux modèles sont bien adaptés à nos données.

Copule	Joe	Gumbel
Statistique	0,41321	0,57224
p-value	0,005	0,005

FIGURE 15 – Test d'ajustement de Cramer-Von Mises

Les données présentées dans le tableau ci-dessus confirment cette conclusion. Même si les copules de Joe et de Gumbel démontrent toutes deux un ajustement satisfaisant, la statistique associée à la copule de Joe est inférieure à celle de Gumbel. Ce détail indique que la copule de Joe présente un meilleur ajustement global à nos données.

En conclusion, nos résultats nous conduisent à privilégier la copule de Joe avec un paramètre estimé à 1,219. Cette copule est retenue comme étant la plus appropriée pour représenter la structure de dépendance entre les variables examinées, offrant ainsi une compréhension précise et détaillée de la relation entre *danceability* et *speechiness* dans les titres musicaux des années 2010.

III Estimation non-paramétrique

À présent, ayant exploré et obtenu des résultats significatifs à travers l'approche paramétrique, tournons notre attention vers l'application d'une méthode non-paramétrique. Cette approche diffère de la précédente du fait qu'elle ne repose pas sur des hypothèses spécifiques concernant la forme fonctionnelle des distributions ou des modèles de copules. Au lieu de cela, la méthode non-paramétrique permet une analyse plus flexible et moins contrainte par des structures prédéfinies.

L'utilisation de méthodes non-paramétriques est particulièrement utile lorsque la nature des données est complexe ou lorsque les relations entre les variables ne se prêtent pas facilement à des modélisations paramétriques standard. Dans le contexte de notre étude, cette approche peut nous fournir des insights supplémentaires sur la dépendance entre *Danceability* et *Speechiness*, en explorant des modèles qui peuvent capturer des caractéristiques subtiles ou non standard dans les données.

En utilisant des techniques non-paramétriques, nous pourrons évaluer la dépendance entre les variables d'une manière qui ne repose pas sur des hypothèses rigides. Cela peut inclure l'utilisation de méthodes de lissage, de tests de dépendance basés sur les rangs, ou d'autres outils statistiques qui ne nécessitent pas de spécifier à l'avance la forme de la distribution ou de la copule.

L'exploration de ces méthodes non-paramétriques nous permettra de comparer et de compléter les résultats obtenus par l'approche paramétrique, offrant ainsi une vue plus complète et nuancée de la structure de dépendance entre nos variables d'intérêt.

III.1 Estimation des paramètres de la copule

Grâce au package **kdecopula** disponible dans R, développé par Nagler, nous avons accès à des outils puissants pour la modélisation non-paramétrique des copules. Ce package fournit des fonctionnalités pour estimer les copules à l'aide d'estimateurs à noyau, une méthode non-paramétrique populaire pour estimer les densités de copules à partir de données échantillonées.

Comme abordé en cours, il existe une variété d'estimateurs à noyau pour estimer les copules d'un échantillon. Ces estimateurs diffèrent dans leur construction et leurs propriétés, et certains sont plus performants que d'autres, notamment en termes de gestion des problèmes liés aux bords des distributions.

Dans notre analyse, nous avons opté pour l'utilisation d'un estimateur à noyau particulier fourni par kdecopula. Ce choix est motivé par le désir d'éviter les problèmes communs rencontrés aux bords de la distribution, qui peuvent fausser les estimations de la copule, surtout dans les zones de faible densité de données ou près des limites de l'échelle des variables.

L'utilisation de cet estimateur à noyau spécifique nous permet d'obtenir une esti-

mation plus fiable et précise de la densité de la copule pour notre échantillon. Cela nous aide à mieux comprendre la structure de dépendance sous-jacente entre *Danceability* et *Speechiness*, sans les contraintes imposées par les modèles paramétriques. Cette approche complète nos résultats paramétriques et offre une perspective plus large sur la relation entre ces deux variables importantes dans les données de musique.

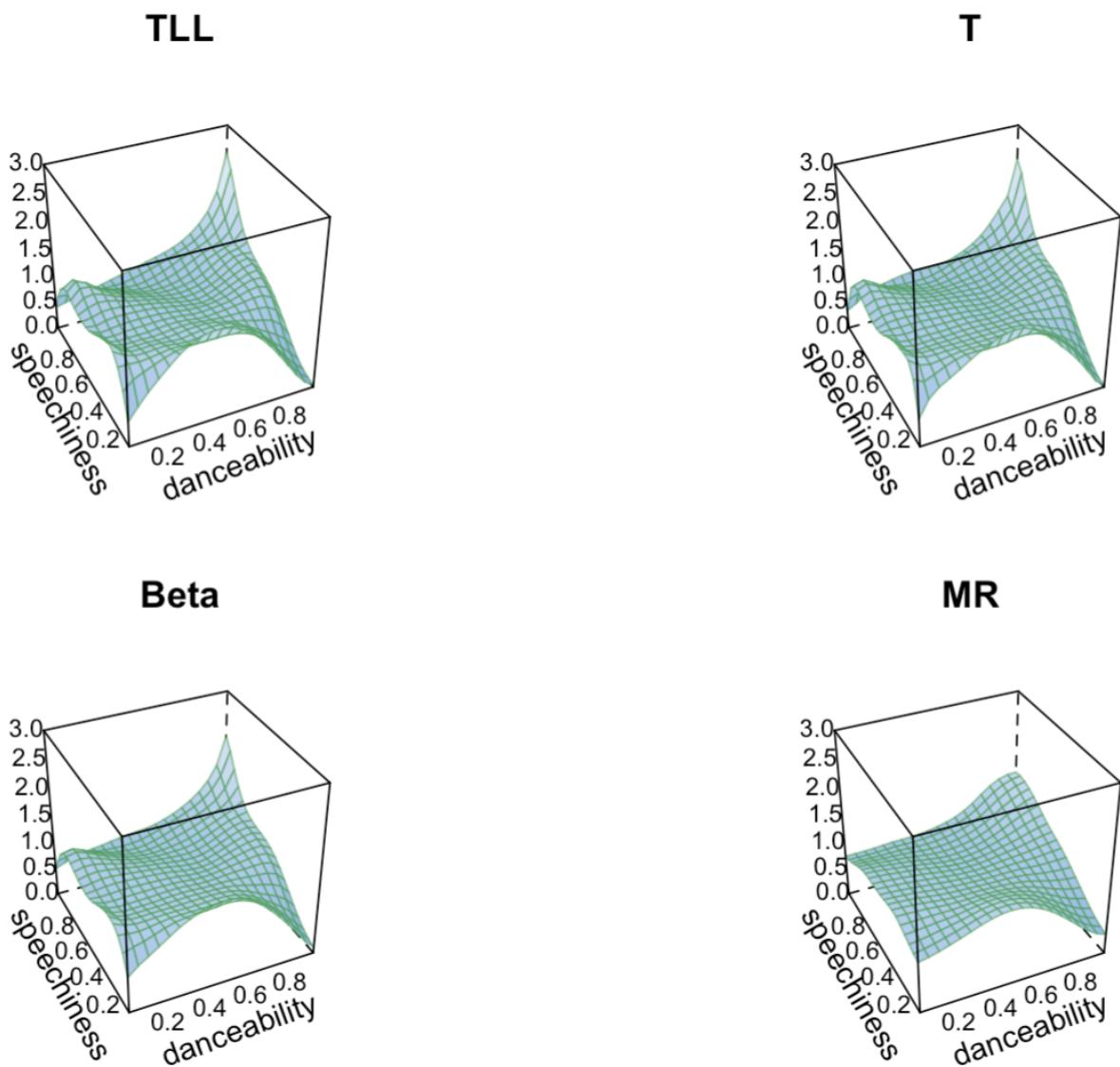


FIGURE 16 – Densités non-paramétriques selon diverses méthodes

Les quatre méthodes que nous avons utilisées pour l'estimation non-paramétrique des densités de copule dans notre étude sont les suivantes :

- **TLL**² : Cette méthode, expliquée par G. Geenens et al. en 2014³, repose sur une transformation locale des données pour maximiser la vraisemblance. La méthode TLL a plusieurs variantes, offrant une flexibilité dans la modélisation de la densité de la copule.
2. Transformation Locale du Maximum de Vraisemblance
 3. Probit Transformation for Non-Parametric Kernel Estimation of the Copula Density

- **T⁴** : Présenté par A. Charpentier et al. en 2006⁵, cet estimateur utilise une transformation des données pour estimer la densité de la copule. L'approche se concentre sur la modification des marges pour obtenir une meilleure estimation de la copule.
- **Bêta** : Cette méthode utilise le Bêta Kernel, comme décrit par A. Charpentier en 2006. L'approche du Bêta Kernel est particulièrement utile pour estimer les densités de copule dans les zones proches des bords, où d'autres méthodes peuvent rencontrer des difficultés.
- **MR⁶** : Développé par I. Gijbels en 1990⁷, le mirror-reflection estimator est une technique qui améliore l'estimation de la densité de la copule en réfléchissant les données autour des bords. Cette méthode aide à réduire les biais et les problèmes liés aux estimations près des limites de la distribution.

Chacune de ces méthodes apporte une perspective unique à l'estimation de la densité de la copule et a été choisie pour ses particularités et ses avantages dans le contexte de notre analyse. L'utilisation de ces différentes approches nous permet d'obtenir une compréhension plus complète et précise de la structure de dépendance entre les variables d'intérêt.

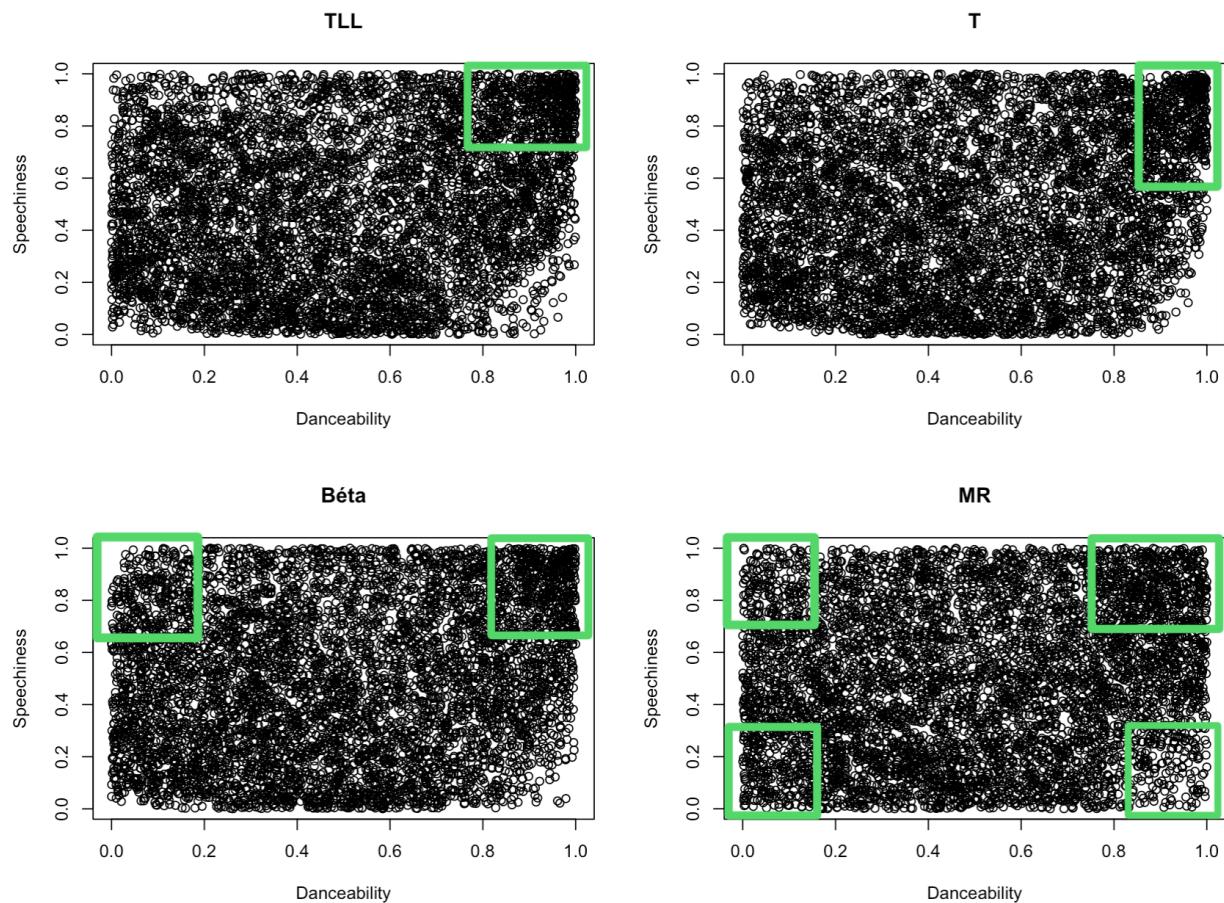


FIGURE 17 – Rank-Rank plot non-paramétrique selon diverses méthodes

4. Estimateur de Transformation
5. The Estimation of Copulas : Theory and Practice
6. Mirror-Reflection Estimator
7. Estimating the Density of a Copula Function

Ces différentes méthodes d'estimation non-paramétrique de la densité de la copule nous ont permis de générer les résultats illustrés dans les FIGURES 16 et 17. Ces résultats sont en adéquation avec ceux présentés dans la première partie de notre étude. En particulier, les densités obtenues sont similaires aux copules de Joe et de Gumbel, ce qui renforce notre choix initial de ces copules classiques pour modéliser la dépendance entre les variables.

Les densités estimées par les quatre méthodes (TLL, T, Bêta et MR) montrent une forte dépendance dans les extrêmes à droite, indiquant une corrélation marquée dans les valeurs élevées des deux variables. De plus, les formes de densité sont similaires entre les méthodes, suggérant une cohérence dans la capture de la structure de dépendance.

Cependant, des différences notables apparaissent lorsqu'on examine les bords des distributions. En se basant sur les rank-rank plots de la FIGURE 17, il apparaît que les méthodes Bêta et MR sont moins performantes, particulièrement en termes d'estimation aux bords. Cette observation est mise en évidence par les cadres verts dans les graphiques et par une concentration moins marquée dans le coin supérieur droit. Ces différences indiquent que, bien que les méthodes Bêta et MR soient utiles dans certains contextes, elles pourraient ne pas fournir l'estimation la plus précise de la dépendance aux bords de la distribution.

En conclusion, ces analyses non-paramétriques complètent et confirment les résultats obtenus par les méthodes paramétriques, tout en offrant des perspectives supplémentaires sur la nature de la dépendance entre nos variables d'intérêt. Cette approche globale renforce notre compréhension de la relation entre *danceability* et *speechiness* dans les données musicales. De plus, il est clair que le choix va principalement devoir se faire entre la méthode T et TLL.

III.2 Qualité de l'ajustement non-paramétrique

Afin de choisir la copule qui nous paraît la plus adaptée à notre échantillon de données, nous avons décidé d'utiliser les critères présentés dans le tableau ci-dessous.

Méthode utilisée	Transformation local likelihood (TLL)	Transformation estimator (T)	Beta Kernel (Beta)	Mirror reflection (MR)
AIC	-895,49	-1 020,46	-789,39	-544,76
BIC	-675,11	-986,25	-540,63	-522,35
Log vraisemblance	480,32	515,29	431,47	275,69

FIGURE 18 – Critères de sélection de la copule non-paramétrique

La méthode T, qui utilise l'estimateur de transformation, se distingue nettement par sa capacité à minimiser les critères AIC (Akaike Information Criterion) et BIC (Bayesian

Information Criterion), tout en maximisant la log-vraisemblance. Cette performance est cohérente avec les observations que nous avons faites précédemment en nous basant sur les densités et les rank-rank plots. Ces résultats indiquent que la méthode T fournit une estimation particulièrement efficace et précise de la densité de la copule pour nos données.

En conséquence, nous choisissons la méthode T comme la méthode la plus appropriée pour l'estimation non-paramétrique de la copule dans notre étude. Cette décision est fondée sur la robustesse et la cohérence des résultats obtenus avec cette méthode.

Conclusion : La meilleure copule non-paramétrique pour notre analyse est celle estimée par l'estimateur de transformation. Cette conclusion est soutenue par des critères statistiques rigoureux et par une analyse visuelle des données, confirmant que cette approche fournit une représentation fidèle et précise de la structure de dépendance entre nos variables d'intérêt.

Conclusion - Choix de la copule

En comparant les résultats obtenus à travers les approches paramétrique et non-paramétrique, notre objectif est de déterminer la copule qui modélise le mieux la dépendance entre la capacité à danser sur une musique des années 2010 et la quantité de paroles dans ces mêmes titres musicaux.

Les analyses paramétriques ont montré que les copules de Joe et de Gumbel sont bien adaptées à nos données, comme en témoignent les p-values favorables et la bonne adéquation globale (*goodness-of-fit*). Ces résultats indiquent que ces copules classiques capturent efficacement la structure de dépendance observée.

D'autre part, l'approche non-paramétrique, en particulier la méthode T utilisant l'estimateur de transformation, a également donné des résultats convaincants. Cette méthode se distingue par des critères AIC et BIC plus faibles et une log-vraisemblance plus élevée, ce qui indique un ajustement supérieur aux données par rapport aux méthodes paramétriques.

Bien que les deux approches fournissent des résultats convaincants, si nous devons faire un choix entre elles, nous privilégierions l'approche non-paramétrique. La raison principale de ce choix est la performance supérieure de la méthode non-paramétrique en termes d'AIC-BIC et de log-vraisemblance, ce qui suggère une meilleure adéquation et une représentation plus précise de la dépendance entre les variables.

En conclusion, notre étude nous conduit à sélectionner la copule non-paramétrique estimée par l'estimateur de transformation comme étant la mieux adaptée pour modéliser la relation entre la capacité à danser sur une musique et la quantité de paroles dans les titres musicaux des années 2010. Cette copule non-paramétrique offre une compréhension approfondie et nuancée de la structure de dépendance, alignée sur notre problématique initiale.

Bibliographie

biCopulaGOF. (s.d.). *Goodness of fit for Bidimensional Copula with Known Margins*. https://genomaths.github.io/usefr_manual/bicopulaGOF.html

cran.r, P. (s.d.). *Copula Modeling*. <https://cran.r-project.org/web/packages/univariateML/vignettes/copula.html>

fitCopula, R. (s.d.). *Fitting Copulas To Data – Copula Parameter Estimation*. <https://www.rdocumentation.org/packages/copula/versions/1.0-1/topics/fitCopula>

Nagler, T. (s.d.). *kdecopula : An R Package for the Kernel Estimation of Copula Densities*. <https://cran.microsoft.com/snapshot/2017-04-02/web/packages/kdecopula/vignettes/kdecopula.pdf>

Annexes

Copule de Joe : Une copule de Joe avec un paramètre de dépendance $\alpha \in [0, \infty[$ a pour générateur la fonction :

$$\forall t \in [0, \infty[, \psi_\alpha(t) = 1 - (1 - e^{-t})^{\frac{1}{\alpha}}$$

Graphiquement, on représente cette copulation avec la densité bivariée suivante, ainsi que le nuage de points (pour $n = 100$) correspondant :

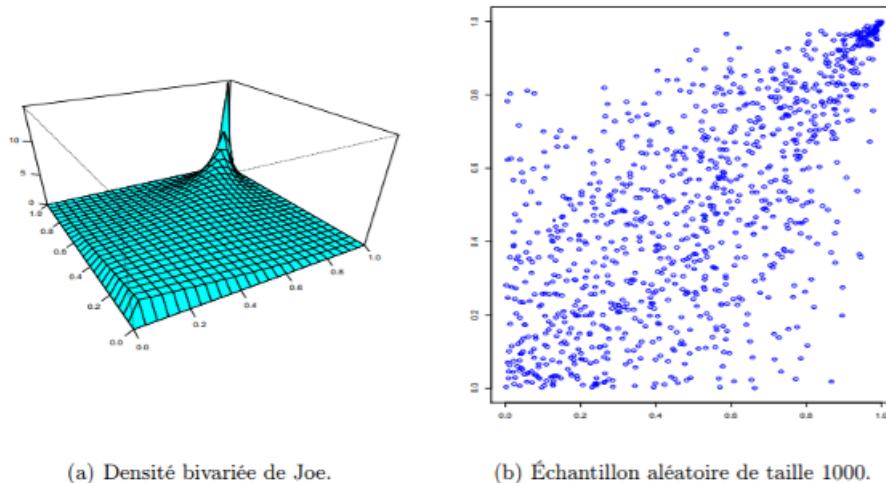


FIGURE 19

On constate sur cette copule, une forte dépendance des extrêmes à droite. De plus, celle-ci peut être comparée à une copule de Gumbel par leurs nombreuses ressemblances en raison de leurs densités similaires.

Loi Kumaraswamy : Il s'agit d'une loi continue dont le support est sur $[0, 1]$ et dépendant de deux paramètres a et b .

C'est une loi similaire à une loi bêta. Sa fonction de densité est donnée par :

$$\forall x \in [0, 1], f(x) = abx^{a-1}(1-x^a)^{b-1}$$



Fin du Rapport