Modélisation des risques en présence de valeurs extrêmes : Une approche Gini

Téa Ouraga

. École doctorale n° 583 : Risques et société

Doctorat de Sciences Économiques

THÈSE

pour obtenir le grade de docteur délivré par

l'Université de Nîmes
EA Chrome

Spécialité doctorale : Économie financière

présentée et soutenue publiquement par

Téa OURAGA

le 07 décembre 2020

Modélisation des risques en présence de valeurs extrêmes : Une approche Gini

Directeur de thèse : **Stéphane Mussard**

Jury

M. Stéphane MUSSARD	Professeur	Université de Nîmes	Directeur de thèse
M. Walter BRIEC	Professeur	Université de Perpignan	Rapporteur
M. Jules SADEFO-KAMDEM	Professeur	Université de Montpellier	Rapporteur
M. Chafic MERHY	Directeur de recherche	Ostrum asset management	Examinateur
M. Jean-Luc PRIGENT	Professeur	Université de Cergy-Pontoise	Examinateur

Université de Nîmes
EA7352 - Détection, Évaluation, Gestion de Risques CHROniques et éMErgents
Rue du Dr G. Salan 30021 Nîmes Cedex 01, France

THESE

REMERCIEMENTS

Je souhaite tout d'abord chaleureusement remercier mon directeur de thèse Monsieur le Professeur Stéphane Mussard à qui j'exprime toute ma gratitude et ma reconnaissance pour la confiance accordée en me prenant sous sa direction pour cette thèse. Ses conseils, ses remarques ont guidé mes premiers pas dans la recherche depuis le Master. Sa qualité humaine et sa patience m'ont permis de m'améliorer sur le plan humain. Je tenais à remercier également sa petite famille à qui je l'ai privé très souvent.

J'adresse mes remerciements les plus sincères à Messieurs les Professeurs Walter Briec (Université de Perpignan) et Jules Sadefo-Kamdem (Université de Montpellier) de m'avoir fait l'honneur d'accepter d'être les rapporteurs de mon travail de thèse. Je souhaite remercier vivement mes examinateurs Jean-Luc Prigent (Université de Cergy-Pontoise) et Chafic Merhy (Ostrum asset management) pour leur disponibilité et cet honneur qu'ils me font.

Mes remerciements vont également à Monsieur le Professeur Arthur Charpentier (Université du Québec à Montréal) pour sa précieuse collaboration qui m'a permis de bénéficier de sa connaissance approfondie en analyse des données et qui a donné lieu à un article en révision.
De même, mes remerciements et ma reconnaissance vont à Madame Françoise Seyte, Maître de Conférence HDR et responsable du Master Finance de marché (Université de Montpellier) qui m'a orienté vers Monsieur Stéphane Mussard et m'a prodigué de sages conseils, utiles durant toutes ces

années de thèse.

Je voudrais remercier les responsables de l'école doctorale *Risques et société*, les responsables du laboratoire *Chrome* qui m'a accueilli durant toutes ses années de thèse. Mes remerciements vont particulièrement à Carine Moulin-Farina, Benoît Roig, Véronique Thireau, Julie Olivero, Karine Weiss et Guillaume Zambrano.
Mes remerciements vont également aux responsables du Groupe de Recherche Angevin en Économie et Management (GRANEM) qui m'ont accueilli durant mes années d'Attaché Temporaire d'Enseignement et de Recherche.

Durant ces années de thèse, j'ai rencontré des collègues qui sont ensuite devenus des amis. La bonne humeur et parfois leurs bonnes idées m'ont permis de passer de belles années de thèse. Je pense à Charlène, Guillaume, Charles, Clément et Julien. Olivier, Aurore, Margaux, Etienne, Arnaud, Waed, Chris, Guillaume, Herell et Runsheng ont facilité mon intégration au sein du GRANEM et m'ont particulièrement soutenu durant la phase de rédaction de mon manuscrit.

Je ne pourrais m'empêcher de chaleureusement remercier mes compagnons de tous les jours et cela commence par mon devancier Alfred, nana Mbaï. J'ai longuement bénéficier de ses conseils durant ces années de thèse et de ses connaissances en théorie financière et en Statistiques. Celui à qui ce projet de thèse doit beaucoup. Merci à Monsieur Gervais Mefoo pour ses conseils avisés. Je remercie également Valé-Manuel pour son partage du savoir et sa présence. Dr Nyamien Yves, Dr Amoussou Landry, Innocent, Malika, Raine, William Guidy, Geraud, Anselme, Larry, Willerm, Sossié, Colombe, Moustapha, Henry-Antoine, Ibrahim, Evans, Cissy, Damaris, Raphael, Aphelly, Herman, Ivan, Franck-Herman, Olivier, Hans Alvin, Dizo, Franklin, Don Habib, Maurine, Alice, Guy-Roland, Alou Jaher et Djê qui m'ont apporté de la bonne humeur. Pour la présence et le soutien sans failles, je remercie sincèrement Ghislaine Kouamé, Marie-Philippe Atsain, Landry Vy-Légré et Aboulaye Diop. Merci également à Marie-Victoire pour sa présence et son aide, mais aussi à ses parents. Mes compagnons de Limoges qui ont vu mes premiers pas en France, peines et joies, je vous remercie affectueusement : Stelvio, Laetitia et Éric.

Enfin mes pensées vont vers ma famille. Je voudrais du plus profond

du coeur remercier mes parents Ouraga Daignekpo Mathurin et Aliman Tchimon Claudine. Ils m'ont inspiré et inculqué des valeurs telles que la persévérance, le travail bien fait et la recherche de l'excellence. Je les remercie d'avoir donné le meilleur d'eux-mêmes, d'avoir renoncer la plupart du temps à eux et de m'avoir donné énormément d'amour pour me mettre dans les meilleures conditions d'études. Ce travail est dédié à feu Yodé Claude, feu Aliman Louis, à ma mère et à mon fils Nolan Ouraga, qui fera sûrement mieux que moi. J'ai aussi une pensée pour mes oncles et tantes qui ont toujours su faire ressortir le meilleur de moi-même. Je témoigne également ma gratitude à mes magnifiques frères et soeurs ; Ahissey, Marie-Paule, Robertine, Annick, Jean-Emmanuel et Andréa qui ont toujours répondu quand j'avais besoin d'eux, mais aussi au patriarche Afran Akomian Vincent et à tous mes cousins, cousines, neveux et nièces. Qu'il me soit permis de partager en cette fin des remerciements la devise du lycée qui m'a donné les fondamentaux et à qui je dois beaucoup, l'École Militaire Préparatoire Technique (EMPT) de Bingerville : **S'instruire pour mieux servir**!

SOMMAIRE

RÉSUMÉ

Cette thèse propose une nouvelle approche du traitement des informations financières disponibles par des outils d'analyse des données robustes, à l'aide de l'indice de Gini. Elle vise à garder toute l'information disponible, même les évènements rares pour la modélisation des rendements et du risque. Une mauvaise appréciation du risque par l'agent fausse ses anticipations. La prise de décision relève de l'évaluation des risques et des prévisions que les agents sont capables de réaliser dans un futur immédiat. Le rendement anticipé de l'investissement par l'agent sera fonction de plusieurs sources de risques selon le modèle d'évaluation par arbitrage. Yitzhaki et Schechtman (2013) ont posé les bases d'une économétrie nouvelle basée sur l'indice de Gini. Ils proposent d'utiliser l'opérateur coGini plutôt que la covariance afin d'étudier des échantillons dont la loi de distribution sous-jacente peut être une loi de distribution autre que la loi normale.

Cette thèse comporte deux apports principaux. Le premier porte sur les méthodes d'analyse des données traditionnelles : ACP, AFD et scoring. Ces méthodes sont adaptées aux valeurs extrêmes par l'utilisation de l'opérateur coGini (ou covariance au sens de gini) : ACP-Gini et AFD-Gini. Le second porte sur l'analyse du couple risques / rendements. Des applications sont faites en théorie financière notamment le pricing des actifs financiers par le modèle d'évaluation par arbitrage et l'analyse des performances des stratégies d'investissement. Au-delà de la finance, les outils développés dans cette thèse peuvent s'appliquer à toute évaluation de risque (risque

climatique, risque de gouvernance, risque lié à l'évaluation d'évènements rares tels que les séismes, le coronavirus, etc.).

ABSTRACT

This thesis proposes a new approach to the treatment of available financial information by robust data analysis tools, using the Gini index. It aims at keeping all the information available, even rare events, for the modeling of returns and risk. A bad appreciation of risk by the agent distorts his expectations. Decision-making is based on the assessment of risks and forecasts that agents are able to make in the immediate future. The expected return on investment by the agent will depend on several sources of risk according to the arbitrage pricing theory. Yitzhaki and Schechtman (2013) have laid the foundations for a new econometrics based on the Gini index. They propose to use the coGini operator rather than the covariance to study samples whose underlying statistic distribution may be a statistic distribution different from the normal distribution.

This thesis has two main contributions. The first deals with traditional data analysis methods : PCA, LDA and scoring. These methods are adapted to extreme values by using the coGini operator (or covariance in the Gini sense) : Gini-PCA and Gini-LDA. The second concerns the analysis of the returns / risk couple. Applications are made in financial theory, in particular the pricing of financial assets by the arbitrage pricing theory and the performance analysis of investment strategies. Beyond finance, the tools developed in this thesis can be applied to any risk assessment (climate risk, governance risk, risk related to the evaluation of rare events such as earthquakes, coronavirus, etc.).

INTRODUCTION GÉNÉRALE

La notion de risque peut être définie de différentes manières. " Le risque [1] est défini comme étant la possibilité, la probabilité d'un fait, d'un évènement considéré comme un mal ou un dommage". Plus simplement, le risque est la possibilité que survienne un évènement indésirable. Théoriquement, il est la probabilité d'occurrence d'un péril probable ou d'un aléa. La notion de risque n'est pas nouvelle et reste au coeur de toute discipline, en particulier l'économie de la finance de marchés.

L'économie étudie l'allocation optimale de ressources rares face à des besoins illimités. La finance consiste quant à elle à allouer l'épargne (ou un financement) nécessaire à la réalisation d'un projet ou d'une opération financière via les marchés de capitaux ou le marché monétaire. Ces marchés ont donc pour principal rôle de mettre en relation les agents économiques à capacité de financement (excédents de financement) et les agents à besoin de financement. On parle de finance directe, lorsque les agents à capacité de financement, financent directement les agents ayant un besoin de financement pour réaliser une opération d'achat ou de vente. Les marchés financiers sont au coeur de ce mode de financement. Une finance dite indirecte représente l'interaction entre les établissements financiers (banques, sociétés de microfinance, etc.) et les agents qui y jouent un rôle majeur. Les établissements et institutions financiers émettent des titres financiers détenus par des épargnants (généralement les ménages) ou collectent des fonds par des dépôts bancaires ou des livrets pour financer les agents économiques ayant des besoins de financement (en général des entreprises, des banques, etc.). On parle aussi de financement par crédit bancaire. En fonction du caractère plus ou moins libéral des marchés financiers, une économie sera donc qualifiée d'économie d'endettement ou d'économie de marchés financiers. Il s'agira d'économie d'endettement lorsque le mode de financement qui prédomine le financement des agents ayant des besoins de financement est le crédit bancaire. Les marchés financiers dans ce cas de figure sont très peu développés.

Une proposition de fusion de ces deux dimensions (économie et finance) donne naissance à l'économie financière. Elle a pour objectif d'obtenir des ressources de financement sous la forme d'emprunts, d'épargnes, etc. pour les allouer, en prenant en compte **des facteurs de risque**. Certains individus révèlent une aversion au risque ou au contraire **recherchent le risque pour**

1. Le dictionnaire LAROUSSE.

augmenter leur espérance de gain. Plusieurs notions de risque peuvent être recensées. Citons-en quelques unes :

— **Le risque de taux d'intérêt** : Il s'agit d'un risque lié à la fluctuation des taux d'intérêt. En dépit du respect des engagements pris par l'émetteur des titres financiers, la fluctuation des taux d'intérêt expose le détenteur de ces titres au risque de moins-value en capital.

— **Le risque de liquidité** : Le risque de liquidité peut être soit le risque lié au manque de liquidité pour honorer une créance, soit le risque qui concerne des placements financiers difficiles à liquider (à vendre).

— **Le risque politique** : Il s'agit un risque qui est lié à une décision du pouvoir politique en place ou à une situation politique : crises socio-politiques, nationalisation ou privatisation d'entreprises, etc.

— **Le risque réglementaire** : Il s'agit du risque de voir un secteur économique (bancaire, pharmaceutique, etc.) affecté par un changement de loi ou de réglementation.

Cette liste ne saurait être exhaustive. La littérature financière tendrait à les classer en deux classes :

1. **Les risques économiques** (risque d'inflation, risques naturels, politiques, etc.) sont issus du monde économique ou la sphère réelle et menacent les flux financiers liés à un titre.

2. **Les risques financiers** (risque de taux d'intérêt, risque de taux de change, risque de liquidité, etc.) sont ceux qui proviennent de la sphère monétaire et financière : évènements financiers externes non imputables à l'entreprise.

L'appréciation de l'impact du risque sur la prise de décision passe inéluctablement par une connaissance et une analyse fine du risque mais aussi par la modélisation de celui-ci. La modélisation peut-être définie comme une représentation synthétique et/ou schématique de la réalité. Elle a pour objectif de prendre des décisions à partir de l'observation d'un ensemble de données. La mesure de ces données peut présenter des

anomalies (erreurs de mesure). De plus, la notion de risque en finance ou en économie financière est assimilée à la dispersion des rentabilités possibles autour de la moyenne des rentabilités. Cette notion implique par exemple la présence de valeurs extrêmes, d'outliers [2], d'évènements rares dans les données.

Comment modéliser des rendements en gardant toute l'information disponible, même les évènements rares ?

La présence de valeurs extrêmes, d'outliers, d'évènements rares dans les données rend inefficace et non convergente la modélisation par la méthode des moindres carrés ordinaires (MCO), basée sur le célèbre théorème de Gauss-Markov. Les estimateurs de cette méthode restent sensibles à la présence de ces derniers. La recherche d'estimateurs peu sensibles ou robustes aux outliers, aux valeurs extrêmes, aux erreurs de mesure s'avère indispensable afin d'obtenir des estimations et prévisions crédibles des rendements des actifs financiers. Cette nécessité a poussé à développer des méthodes robustes telles que la régression médiane, la régression quantile, la régression floue, la régression Gini, pour n'en citer que quelques exemples.

La régression Gini en plus d'être robuste aux outliers, aux valeurs extrêmes, aux erreurs de mesure, est équivalente aux MCO [3] lorsque les hypothèses de normalité tiennent, et est plus robuste lorsque les hypothèses sont violées. Yitzakhi, Schechtman et Pudalov en 2011 proposent deux régressions qui peuvent être interprétées comme basées sur la différence moyenne de Gini (GMD [4]) : une approche semi-paramétrique, qui repose sur la moyenne pondérée des pentes définies entre des observations adjacentes et une approche de minimisation, qui repose sur la minimisation de la GMD des résidus. Les estimateurs obtenus par l'approche semi-paramétrique ont des représentations qui ressemblent aux estimateurs MCO. En outre, ils sont plus robustes que les estimateurs MCO aux

2. Outliers ou valeurs aberrantes.

3. Elle consiste à minimiser la somme des carrés des écarts entre chaque point du nuage de régression et son projeté sur la droite de régression (erreurs ou résidus).

4. La GMD est un indice de variabilité alternatif à la variance, qui est l'indice homogène de degré 1(alors que l'indice de Gini standard est homogène de degré 0). Elle partage de nombreuses propriétés avec la variance, mais peut être plus informative sur les propriétés des distributions statistiques qui s'écartent de la normalité.

observations extrêmes et insensibles aux transformations monotones. De même Ka et Mussard (2015) proposent une régression sur norme ℓ_1 pour les données de panel. Il est montré dans cet article que l'estimateur Gini à effets fixes au sein d'un groupe est plus robuste que l'estimateur des moindres carrés ordinaires lorsque les données sont contaminées par des observations aberrantes.

Pour les décideurs la gestion de l'information financière est un enjeu important pour la gestion de risque et pour faire face au besoin de prédictibilité de la performance financière. La modélisation financière reste une des solutions les plus utilisées qui permet de faire face à cette problématique. Elle peut se définir comme le processus par lequel des projections de flux financiers sont réalisées par un agent économique en fonction de divers scénarii. La modélisation financière permet donc d'élaborer des modèles financiers, tels que le modèle d'évaluation des actifs financiers (MEDAF ou CAPM en anglais), développé au début des années 1960 à partir des travaux de Markowitz, Sharpe, Lintner et Treynor, qui permettent de développer des instruments d'analyse et de simulation des situations (ou scénarii) qui contribuent à prendre des décisions parfaitement adaptées. Ces schématisations simplifiées (ou représentations) d'une situation financière s'appuient sur des outils mathématiques ou économétriques dont la fiabilité des estimateurs est essentielle à la bonne prise de décision en prenant en compte un certain nombre de facteurs. La modélisation financière vise trois objectifs :

1. Premièrement, à travers la mise en oeuvre de projections, la modélisation financière facilite la prise de décisions car elle donne les moyens pour définir le niveau de sensibilité des variables les plus utiles.
2. Deuxièmement, la modélisation financière minimise de manière considérable le degré d'incertitude lorsqu'elle est correcte.
3. Troisièmement, elle permet in fine d'évaluer l'efficacité de la prise de décision grâce à la sensibilité de l'objet de l'étude (la tarification d'un actif financier par exemple) aux principaux indicateurs financiers et une appréciation fiable du niveau de risque lié aux variables les plus importantes.

C'est dans cet élan que les analystes quantitatifs et les chercheurs utiliseront les méthodes robustes de modélisation pour la modélisation financière

dont la modélisation Gini en finance afin d'estimer avec plus de précision le risque. Shalit et Yitzhaki (1984) proposent une tarification des actifs financiers risqués et la construction de portefeuilles d'actifs optimaux en théorie du portefeuille avec une approche moyenne-Gini. Dans cet article, il est montré que le modèle respecte les caractéristiques de dominance stochastique et est conforme au comportement des investisseurs dans l'incertitude pour une large classe de distributions statistiques. La GMD est plus adéquate que la variance pour mesurer la variabilité et donc le risque d'un actif. Ce modèle est également étendu pour inclure un degré d'aversion au risque qui peut être estimé à partir de données réelles d'un marché financier grâce au GMD généralisé. Parallèlement en 2001, les mêmes auteurs prouvent que les estimateurs des moindres carrés ordinaires des coefficients bêta des grandes entreprises sont très sensibles aux observations extrêmes des rendements des indices de marché. Cette sensibilité provient de la fonction quadratique utilisée (la variance) communément en théorie financière. Shalit et Yitzhaki en introduisant des considérations d'aversion au risque dans la procédure d'estimation, utilisant des estimateurs alternatifs dérivés des mesures de Gini de la variabilité, montrent qu'il est possible de surmonter ce manque de robustesse et d'améliorer la fiabilité des résultats.
Mussard et Terraza (2004) ont une nouvelle approche de l'estimation des risques par l'indice de Gini, qui est appliquée à la décomposition de la volatilité d'un portefeuille d'actifs financiers. En effet, ils essaient d'établir à travers cet article, un lien entre le risque financier et la classe des mesures d'inégalités décomposables. Cette approche a permis, suite à une application sur des actifs du marché financier français, de construire des indicateurs de risque. Les résultats poussent à remettre en question la nature et la définition de la notion de risque en finance.

A la fin des années 2000, les données massives (ou Big Data) créent un fort intérêt pour les décideurs, ingénieurs, data analystes, etc. Ceux du monde de la finance ne resteront pas en marge avec la gestion de l'information financière qui prend une place capitale dans la gestion des risques et la prise de décision. La numérisation de l'information (données textuelles par exemple) et l'appétence des décideurs pour ces données massives amènent les sciences informatiques à la création de bases données et de serveurs externes pour accumuler l'information nécessaire à la prise de décision. L'exploration, la visualisation et l'analyse de ces données

massives pour l'extraction d'informations pertinentes ont fait naître et ont permis d'améliorer des algorithmes pour le traitement des données : machine learning (forêts aléatoires, machine à vecteurs de support, etc.), les régressions (régressions régularisées, analyse linéaire discriminante, etc.) et les analyses factorielles (Analyse en Composantes Principales, Analyse des Correspondantes Multiples, etc.).

L'objectif de cette thèse est double. D'une part, elle consiste à traiter les informations financières disponibles par des outils d'analyse de données robustes à l'aide de l'indice de Gini. D'autre part, d'appliquer ces outils à des modèles de la Finance (le modèle d'évaluation par arbitrage et le ratio de Treynor généralisé) aidant à la prise de décision en utilisant une estimation non biaisée du risque et la proposition de nouvelles mesures de performance.

Cette thèse aborde la mise en place d'outils d'analyse de données robustes à la présence d'observations extrêmes et aux erreurs de mesures à l'aide de l'indice de Gini. Notamment l'Analyse en Composantes Principales par une approche Gini (ACP-Gini), *Cf.* Ouraga (2019), et Charpentier, Mussard et Ouraga (2020). Cette technique nécessite une maximisation de la matrice de corrélation au sens de Gini de métrique ℓ_1. De l'ACP-Gini, nous construisons des outils pour la prise de décision en gestion de portefeuille : le Modèle d'Évaluation par Arbitrage et le Ratio de Treynor Généralisé par une approche Gini, respectivement MEA-Gini et RTG-Gini. La décomposition de la matrice de corrélation Gini généralisée en matrice Gini généralisée intra-groupes et matrice Gini généralisée inter-groupes permet de proposer les analyses discriminantes linéaire et quadratique au sens de Gini. Nous montrons à l'aide de simulations de Monte-Carlo la robustesse et la supériorité de cette méthode sur d'autres classifieurs tels que la régression logistique, les SVM et l'analyse linéaire discriminante classique, *Cf.* Condevaux, Mussard, Ouraga & Zambrano (2020).

Pour mener à terme ce double objet de la thèse, dans le Chapitre 1 une revue de la littérature de la régression Gini est proposée afin de mettre en évidence les outils de la méthodologie Gini utilisés dans les chapitres suivants, *Cf.* Ouraga (2018) . Dans le Chapitre 2, nous montrons qu'un des éléments de la méthodologie à savoir la matrice de corrélation au sens de Gini permet d'obtenir une ACP-Gini robuste aux observations extrêmes. Dans le Chapitre 3, nous mettons en place une extension de l'ACP-Gini

ou ACP-Gini généralisée qui fait intervenir la matrice de corrélation généralisée avec un paramètre d'aversion au risque. En se basant sur cet outil robuste aux outliers, le Chapitre 4 propose de revisiter quelques éléments de la gestion de portefeuilles dans un univers moyenne-GMD telle que la frontière efficiente. Cette approche dans un univers moyenne-GMD induit une mesure du risque (ou de la volatilité) par la GMD, une mesure des sensibilités aux facteurs par les estimateurs de la régression Gini. Nous proposons le modèle d'évaluation par arbitrage à facteurs latents en utilisant des facteurs latents issus de l'approche ACP-Gini. De ce modèle sera issu le ratio de Treynor généralisé par une approche Gini (RTG-Gini). Enfin dans le Chapitre 5, une décomposition de la matrice de corrélation Gini est effectuée pour la mise en oeuvre des analyses discriminantes linéaire et quadratique Gini. Des algorithmes sont proposés ainsi qu'une méthode de scoring robuste et plus performante que certains classifieurs tels que la régression logistique ou l'analyse linéaire discriminante classique. Une application aux portefeuilles "verts" du CAC40 sur la période allant du 30/01/2018 au 01/02/2019 confirme ces prédictions théoriques.

CHAPITRE 1

LA RÉGRESSION GINI : UNE REVUE DE LA LITTÉRATURE

Sommaire

Résumé

Ce chapitre propose une revue de la littérature de la méthodologie. La méthodologie Gini, dans le cadre des modèles de régression, a été retenue pour ses performances en cas d'erreur de mesures ou d'outliers venant contaminer les données. La régression Gini met en évidence des estimateurs robustes donnant de meilleurs résultats que les estimateurs LAD ou MCO lorsque les outliers sont présents dans les régresseurs.

Mots-clés : Erreurs de mesure ; Outliers ; Régression Gini ; Robustesse

Abstract

This chapter provides a reivew of the Gini's methodology. The Gini's methodology, within the framework of the models of regression, was retained for its performances in cases of error of measures or outliers coming to contaminate the data. The regression Gini highlights strong estimators giving better results that LAD or MCO estimators when outliers are present in regressors.

Keywords : Measurement errors ; Outliers ; Gini Regression ; Robustness

1.1 Introduction

La modélisation économétrique est très répandue dans les domaines scientifiques tels que l'économie, la finance, la sociologie, la biologie, etc. La recherche d'estimateurs robustes, peu sensibles aux valeurs extrêmes (outliers) s'avère indispensable afin d'obtenir des estimations et prévisions crédibles permettant de valider les raisonnements et hypothèses issus de modèles théoriques.

La statistique « Gini's Mean Difference » (GMD), ou coefficient de Gini absolu, est une mesure alternative de variabilité introduite par le statisticien italien Corrado Gini en 1912. Elle sera très répandue, dans un premier temps en économie du développement en tant qu'indice d'inégalité, puis progressivement s'étendra vers d'autres domaines tels que la finance en tant qu'indice de dispersion (mesure de risque) et l'économétrie.

Entre 1970 et 1990, de nouvelles méthodes de régressions ont été introduites comme les régressions quantiles de Basset et Koenker (1978), la régression linéaire floue de Tanaka et al. (1982), et la régression Gini de Olkin et Yitzhaki (1992) basée sur la statistique GMD. La régression Gini possède deux particularités : elle permet de s'affranchir de certaines hypothèses standards en économétrie et autorise l'utilisation de données comportant des erreurs de mesures ou des valeurs aberrantes.

La présence d'outliers dans la base de données rend difficile l'utilisation de certaines techniques statistiques et de data mining, car elle donne des résultats fallacieux lorsqu'elle n'est pas correctement traitée en amont. De même, le non-respect des hypothèses de base d'un modèle économétrique comme celui de l'exogénéité des variables explicatives dans la méthode des moindres carrés ordinaires (MCO) conduit à de mauvaises interprétations des coefficients (coefficients non significatifs et parfois avec des signes inversés).

L'objet de ce chapitre est de mettre en évidence, dans le cadre des modèles de régressions généralisés, l'utilisation du GMD qui permet de répondre aux problèmes posés par les outliers et les erreurs de mesure sur les variables explicatives, et de s'affranchir de certaines hypothèses de base en économétrie comme la linéarité, *Cf.* Ouraga (2018). Quand la distribu-

tion du processus générateur suit une loi normale univariée, la moyenne et la variation de l'échantillon sont des statistiques suffisantes pour décrire la distribution, rendant l'utilisation du GMD superflu. De même, dans le cas multivarié, lorsque la distribution des régresseurs suit une loi normale multivariée, l'estimateur issu du GMD est équivalent à celui des MCO. Néanmoins, lorsque la distribution n'est pas normale multivariée, le GMD se révèle être un estimateur robuste. Yitzhaki et Schechtman (2013) montrent que le GMD est utile lorsque les relations entre les variables aléatoires sont symétriques ou non, lorsque les rangs des variables aléatoires sont liés, lorsque la population est stratifiée, lorsque l'hypothèse de linéarité du modèle de régression est soutenue ou non par les données observées. La méthodologie Gini permet à son utilisateur d'éstimer les coefficients d'un modèle de régression en utilisant les rangs des régresseurs, d'utiliser les U-statistiques pour l'inférence des coefficients, de tester la linérité du modèle, et de vérifier si les erreurs de mesure dans les régresseurs peuvent avoir une influence sur les estimations.

Cette revue de la littérature de la régression Gini est structurée comme suit : la première section introduit l'opérateur de Gini covariance basé sur les rangs des variables aléatoires ; la deuxième section décrit les différents types de régression Gini ; la troisième section présente les conséquences des erreurs de mesure sur les estimateurs Gini ; la quatrième section expose l'inférence des estimateurs Gini avec la théorie des U-statistiques ; enfin la dernière section conclut l'article.

1.2 L'opérateur de Gini covariance

Nous exposons dans cette section l'opérateur de Gini covariance introduit par Schechtman et Yitzhaki (1987), la notion de Gini corrélation qui en découle, et l'intérêt d'utiliser cette corrélation issue des rangs des variables aléatoires dans les modèles de régression, comme l'avait indiqué Durbin (1954).

Définissons tout d'abord les notations utilisées :
$\hookrightarrow$ $\mathbf{y}$ le vecteur $N \times 1$ de la variable à expliquer (à valeur dans $\mathbb{R}$).
et d'écart-type empirique sans biais s_y.
$\hookrightarrow$ $\mathbf{X}$ la matrice $n \times K$ des variables explicatives, avec $\mathbf{x}_k$ une variable explicative (un vecteur), $k = 1 \dots K$, $K < n$ et x_{ik} la i ème observation du

vecteur $\mathbf{x}_k$, $i = 1, \ldots, n$.
$\hookrightarrow$ $\mathbf{Z}$ la matrice $n \times K$ des variables instrumentales, avec $\mathbf{z}_k$ une variable instrumentale (un vecteur), $k = 1 \ldots K$, $K < n$ et z_{ik} la i ème observation du vecteur $\mathbf{z}_k$, $i = 1, \ldots, n$.
$\hookrightarrow$ $\widehat{\boldsymbol{\beta}} = \widehat{\beta}_1, \ldots, \widehat{\beta}_K$, le vecteur des coefficients estimés.
$\hookrightarrow$ $\boldsymbol{\theta}$ le vecteur de contamination de dimension $n \times 1$.
$\hookrightarrow$ $\widehat{\beta^{\boldsymbol{\theta}}}_G$, le coefficient estimé dans la régression Gini en présence de l'erreur d'observation.
$\hookrightarrow$ $\widehat{V}$ & $\widehat{\sigma}$ respectivement la variance estimée et l'écart-type estimé.
$\hookrightarrow$ $\rho_{X,Y}$ le coefficient de corrélation de Pearson.
$\hookrightarrow$ $Cov\ (X, Y)$ est la covariance entre les variables X et Y.
$\hookrightarrow$ F_X et D_X désignent respectivement la fonction de répartition de X et l'ensemble de définition de F_X.

Définitions

Le coefficient de Gini, ou indice de Gini, est une mesure homogène de degré zéro qui permet de mesurer des disparités (inégalités) au sein d'une population donnée. Il est compris entre 0 et 1. Pour une population oú il existe une parfaite répartition (des revenus), ce coefficient est égal à 0, et inversement. L'indice de Gini homogène de degré 1, encore appelé « Gini's Mean Difference » ou GMD, est un indice de variabilité d'une variable aléatoire. Il définit la valeur attendue entre deux observations prises au hasard dans une population.

Soient X_1 et X_2 des observations indépendantes issues d'une même variable aléatoire X de fonction de répartition F_X. Le GMD est donné par :

$$G_X = \mathbb{E}|X_1 - X_2| \tag{1.1}$$

Remarquons que la variance peut aussi s'écrire en terme d'écarts absolus espérés, mais avec une métrique différente :

$$\sigma_X^2 = \frac{1}{2}\,\mathbb{E}|X_1 - X_2|^2 \tag{1.2}$$

L'indice GMD peut être récrit à l'aide de l'opérateur de Gini covariance mesurant la variabilité entre la variable aléatoire et sa fonction de répartition (Stuart, 1954) :

$$G_X = 4\ Cov(X, F_X) \tag{1.3}$$

Lorsque X est distribuée selon une loi normale, $G_X = 2\sigma_x/\sqrt{\pi}$, alors l'indice de Gini et la variance deviennent des mesures de dispersion équivalentes, dans la mesure où elles renvoient à un même préordre classant deux alternatives.

L'expression (1.3) met en évidence la covariance au sens de Gini. Pour deux variables aléatoires X et Y, il existe précisément deux Gini covariances ($GCov$), introduites par Schechtman et Yitzhaki (1987) :

$$GCov(Y, X) = Cov\left(Y, F_X\right) \tag{1.4}$$
$$GCov(X, Y) = Cov\left(X, F_Y\right) \tag{1.5}$$

La Gini covariance est la mesure de la covariance entre une variable aléatoire et la fonction de répartition (ou le vecteur rang) d'une autre variable aléatoire. Habituellement pour étudier la relation qui existe entre deux variables aléatoires, on utilise la covariance usuelle $Cov(X, Y)$ et le coefficient de corrélation de Pearson dont la métrique est euclidienne (L_2). Afin de changer de métrique, le coefficient de corrélation de Pearson peut être remplacé par celui de Spearman qui nécessite au préalable que les deux variables aléatoires X et Y soient représentées par leur fonction de répartition, $Cov(F_X, F_Y)$. La métrique obtenue est de type L_1 (distance de Manhattan). La Gini covariance peut donc être considérée comme un mélange entre les deux métriques précédentes, L_2 et L_1 respectivement.

La Gini corrélation

La Gini corrélation, notée GC, est une mesure normalisée de corrélation issue de la Gini covariance. Elle prend ses valeurs dans l'intervalle $[-1; 1]$. Dans la mesure où la Gini covariance n'est pas syémtrique, voir Eq.(1.3), deux Gini corrélations peuvent être définies :

$$GC(Y, X) = \frac{GCov(Y, X)}{GCov(Y, Y)} = \frac{Cov(Y, F_X)}{Cov(Y, F_Y)} \tag{1.6}$$

$$GC(X, Y) = \frac{GCov(X, Y)}{GCov(X, X)} = \frac{Cov(X, F_Y)}{Cov(X, F_X)} \tag{1.7}$$

Proposition 1.2.1. – Schechtman et Yitzhaki (1987) :
(i) *Soit deux variables aléatoires X et Y interchangeables, alors il existe une*

fonction $h : \mathbb{R} \to \mathbb{R}$ *telle que* $Xh(Y) = Yh(X)$ *et donc* $\mathbb{E}(Xh(Y)) = \mathbb{E}(Yh(X))$. *L'interchangeabilité implique que* $GC(X, Y) = GC(Y, X)$.
(ii) *Si* (X, Y) *suit une distribution normale bivariée d'espérance* (μ_X, μ_Y), *de variances* σ_X^2, σ_Y^2, *tel que* $\rho_{X,Y}$, *alors :*

$$GC(X, Y) = GC(Y, X) = \rho_{X,Y} \tag{1.8}$$

(iii) *Soit* X *et* Y *deux variables aléatoires, alors :*

$$GC_{X+Y} = GC(X, X + Y)\, G_X + GC(Y, X + Y)\, G_Y. \tag{1.9}$$

En contraste avec la nature symétrique de l'opérateur usuel de covariance, les Gini corrélations $GCov(X, Y)$ et $GCov(Y, X)$, peuvent être de signe et d'intensité différents. Cette propriété peut être vue *a priori* comme une limite de la méthode Gini. Cependant, comme l'ont démontré Carcea et Serfling (2015) dans le cadre des séries temporelles, le fait de disposer de deux fonction d'autocorrélation de type Gini permet de mieux identifier les processus ARMA (notamment en présence de valeurs aberrantes).

La méthode des rangs de Durbin

Afin de comprendre l'intérêt de l'opérateur de Gini covariance dans le cadre de la modélisation économétrique, revenons sur l'apport de Durbin (1954). Ce dernier propose d'utiliser les rangs des variables explicatives comme instruments. Utilisons la statistique d'ordre $\tilde{\mathbf{x}}$ d'une réalisation de la variable aléatoire X telle que $\tilde{x}_1 \leq \tilde{x}_1 \leq \cdots \leq \tilde{x}_n$. Le rang de l'observation i de la variable observée $\mathbf{x}$ issue d'un échantillon est :

$$\mathbf{r}_{x_i} = \sum_{i=1}^{n} \mathbb{1}(x \leq \tilde{x}_i) \tag{1.10}$$

avec $\mathbb{1}(x \leq \tilde{x}_i)$ la fonction qui retourne la valeur 1 lorsque l'expression $x \leq \tilde{x}_i$ est vraie. Un estimateur basique de la fonction de répartition est ainsi obtenu :

$$\widehat{F}_X = \frac{\mathbf{r}_\mathbf{x}}{n} \tag{1.11}$$

avec $\mathbf{r}_\mathbf{x} = (\mathbf{r}_{x_1}, \ldots, \mathbf{r}_{x_n})$. Yitzhaki et Schechtman (2013) indique que les *ex aequo* peuvent induire des biais dans le calcul de l'estimateur de la Gini covariance. Dans la pratique, trois méthodes sont couramment utilisées pour traiter les *ex aequo* :

• utiliser le rang supérieur des deux observations, ce qui conduira à une sur-pondération et donc un biais positif ;

• utiliser le rang inférieur, ce qui aboutira à une sur-pondération et donc un biais négatif ;

• estimer le rang au point moyen, ce qui produira des estimateurs sans biais.

Par exemple, la troisième méthode peut être illustrée de la manière suivante :

$$\mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 4 \\ 7 \\ 6 \end{pmatrix} \longrightarrow \mathbf{r_x} = \begin{pmatrix} 1,5 \\ 1,5 \\ 3 \\ 5 \\ 4 \end{pmatrix} \tag{1.12}$$

Les estimateurs de la Gini covariance s'expriment donc comme suit :

$$\widehat{GCov}(X,Y) = \frac{1}{n}\, Cov(\mathbf{x}, \mathbf{r_y}) \tag{1.13}$$

$$\widehat{GCov}(Y,X) = \frac{1}{n}\, Cov(\mathbf{y}, \mathbf{r_x}) \tag{1.14}$$

En 1954, Durbin fait remarquer que lorsque des données comportent des valeurs aberrantes, il est utile d'utiliser le vecteur rang dont les valeurs restent relativement stables en cas de contamination. Reprenons l'exemple précédent, en multipliant la plus forte observation par 100, le vecteur rang reste inchangé :

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 4 \\ 700 \\ 6 \end{pmatrix} \longrightarrow \mathbf{r_{x_1}} = \begin{pmatrix} 1,5 \\ 1,5 \\ 3 \\ 5 \\ 4 \end{pmatrix} = \mathbf{r_x} \tag{1.15}$$

Durbin (1954) propose par conséquent d'utiliser la matrice d'instruments $\mathbf{R} = (\mathbf{r_{x_1}}, \ldots, \mathbf{r_{x_K}})$ qui contient en colonne les vecteurs rangs des variables explicatives. Puisqu'un vecteur rang peut raisonnablement être considéré comme indépendant du terme d'erreur, des estimateurs robustes peuvent être obtenus dans le cas d'une régression multiple de type $y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

$$\hat{\boldsymbol{\beta}}_{VI-rank} = (\mathbf{R}'\boldsymbol{X})^{-1}\, \mathbf{R}'\mathbf{y} \tag{1.16}$$

En notant $\mathbf{Z}$ la matrice des instruments, telle que $\boldsymbol{Z} = \mathbf{R}$, on retrouve l'estimateur usuel $(\mathbf{Z}'\boldsymbol{X})^{-1}\,\mathbf{Z}'\mathbf{y}$ des moindres carrés ordinaires par variables instrumentales dans le cas où il existe autant d'instruments que de régresseurs. Durbin (1954) venait, sans le savoir, de découvrir une forme particulière de la régression Gini.

1.3 La régression Gini

La régression Gini compte deux approches. La première approche consiste à minimiser une mesure de dispersion des résidus, alternative à la variance. Il s'agit simplement de changer de norme en utilisant l'indice de Gini des résidus, autrement dit passer de la norme L_2 à la norme L_1. Cette première approche, par minimisation, est l'approche paramétrique. La seconde, l'approche semi-paramétrique, consiste à trouver des estimateurs robustes à partir de moyennes pondérées d'estimateurs de tendance centrale comme la médiane. L'idée est de trouver un système de pondération moins sensible aux outliers aboutissant à des propriétés désirables pour les estimateurs obtenus.

Hypothèses classiques

Revenons sur les hypothèses standards des modèles linéaires usuels $\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$.

Hypothèses 1.3.1. – **(H1)** – : *Le modèle est linéaire en $\mathbf{x}_k$ (ou en n'importe quelle transformation de $\mathbf{x}_k$).*

Hypothèses 1.3.2. – **(H2)** – : *Les valeurs x_{ik} sont observées sans erreur (x_{ik} non aléatoire).*

Hypothèses 1.3.3. – **(H3)** – : $\mathbb{E}(\epsilon_i^2) = \sigma_{\boldsymbol{\epsilon}}^2$, *la variance de l'erreur est constante (homoscédasticité) : le risque de l'amplitude de l'erreur est constante quelle que soit la période.*

Hypothèses 1.3.4. – **(H4)** – : $\mathbb{E}(\epsilon_i \epsilon_{i'}) = 0$ *si $i \neq i'$, les erreurs sont non corrélées ou indépendantes*

Hypothèses 1.3.5. – **(H5)** – : $\mathbb{E}(\mathbf{x}_k' \boldsymbol{\epsilon}) = 0$, *l'erreur est indépendante de la variable explicative k, pour tout $k = 1, \ldots, K$.*

Il est intéressant de mettre en évidence, dans le cadre des régressions Gini, les hypothèses nécessaires à leur mise en oeuvre et celles qui peuvent être relâchées.

La regression Gini paramétrique

La régression paramétrique est basée sur la minimisation du Gini des résidus. Pour procéder par minimisation, la forme du modèle doit être spécifiée. L'hypothèse (H1) est donc invoquée.

Considérons le modèle linéaire simple suivant $\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x} + \boldsymbol{\epsilon}$. Pour un échantillon de taille n, la valeur estimée de $\mathbf{y}$ est notée $\hat{\mathbf{y}} = \hat{\alpha} + \hat{\beta}\mathbf{x}$. L'indice de Gini du résidu $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ est une fonction de $\hat{\beta}$:

$$G_{\mathbf{e}}(\hat{\beta}) = \frac{1}{n} Cov(\mathbf{e}, \mathbf{r}_{\mathbf{e}}), \tag{1.17}$$

où $\mathbf{r}_{\mathbf{e}}$ représente le vecteur rang du résidu. Minimiser (1.17) revient à minimiser $\sum_{i=1}^{n} e_i \mathbf{r}_{e_i}$ qui est la fonction de distance utilisée dans la R-regression de Jurecková (1981), Jaeckel (1972) et McKean & Hettmansperger (1976). L'indice de Gini du résidu se récrit :

$$\begin{aligned} G_{\mathbf{e}}(\hat{\beta}) &= \frac{4}{n} Cov(\mathbf{e}, \mathbf{r}_{\mathbf{e}}) \\ &= \frac{4}{n} Cov(\mathbf{y} - \hat{\alpha} - \hat{\beta}\mathbf{x}, \mathbf{r}_{\mathbf{e}}) \\ &= \frac{4}{n} \Big[Cov(\mathbf{y}, \mathbf{r}_{\mathbf{e}}) - \hat{\beta}\, Cov(\mathbf{x}, \mathbf{r}_{\mathbf{e}}) \Big] \\ &= \frac{4}{n} \Big[Cov(\mathbf{y}, \mathbf{r}_{\mathbf{e}}) - \hat{\beta}\, Cov(\mathbf{x}, \mathbf{r}_{\mathbf{e}}) \Big] \end{aligned} \tag{1.18}$$

Pour un $\hat{\beta}$ donné, calculons l'indice de Gini du résidu en utilisant l'écart espéré entre deux observations prises au hasard (avec remise) :

$$G_{\mathbf{e}}(\hat{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|e_i - e_j|}{n^2} \tag{1.19}$$

Minimiser $G_e(\widehat{\beta})$ revient à minimiser :

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}|e_i - e_j| &= \sum_{i,j}|(y_i - \widehat{y}_i) - (y_j - \widehat{y}_j)| \\ &= \sum_{i,j}|(y_i - y_j) - (\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}x_j)| \\ &= \sum_{i,j}|(y_j - y_i) - \hat{\beta}\,(x_j - x_i)| \\ &= 2\sum_{i<j}[\,(y_j - y_i) - \hat{\beta}\,(x_j - x_i)\,] \qquad (1.20)\end{aligned}$$

Proposition 1.3.1. – Olkin et Yitzhaki (1992) – : *Lorsque l'indice de Gini du résidu est à son minimum,* $G_{\mathbf{e}}(\hat{\beta}) = 2\sum_{i<j}[\,(y_j - y_i) - \hat{\beta}\,(x_j - x_i)\,] = 0$*, la pente de la régression Gini notée* $\hat{\beta}_G$ *est alors :*

$$\widehat{\beta}_G = \frac{\sum_{i<j}(y_j - y_i)}{\sum_{i<j}(x_j - x_i)} \qquad (1.21)$$

Si les résidus sont ordonnés $e_1 \leq e_2 \leq \ldots \leq e_n$, la dérivée de l'indice de Gini en $\hat{\beta}$ est :

$$\begin{aligned}\frac{\partial G_{\mathbf{e}}(\widehat{\beta})}{\partial\widehat{\beta}} &= -2\sum_{i<j}(x_j - x_i) \\ &= 4\sum_{i=1}^{n}x_i\left[i - \frac{n+1}{2}\right] \\ &= 4\,n\,Cov\,(\,\mathbf{x}, \mathbf{r_e}\,) \qquad (1.22)\end{aligned}$$

Par construction, l'indice de Gini atteint son minimum lorsque le terme $Cov(\,\mathbf{x}, \mathbf{r_e}\,)$ est minimisé, ce qui est le cas lorsque $Cov(\,\mathbf{x}, \mathbf{r_e}\,) = 0$.

Proposition 1.3.2. – Olkin et Yitzhaki (1992) – : *Lorsque l'indice de Gini du résidu est minimal, la covariance entre l'estimateur* $\hat{y}$ *et le rang du résidu est nulle :*

$$\begin{aligned}Cov(\hat{y}, \mathbf{r_e}) &= Cov(\alpha + \hat{\beta}_G\mathbf{x}, \mathbf{r_e}) \\ &= \hat{\beta}_G Cov(\mathbf{x}, \mathbf{r_e}) = 0 \qquad (1.23)\end{aligned}$$

Proposition 1.3.3. – Olkin et Yitzhaki (1992) – : *Lorsque l'indice de Gini du résidu est minimal, la covariance entre la variable dépendante* y *et le rang du résidu*

$\mathbf{r_e}$ *est égal à l'indice de Gini du terme d'erreur :*

$$\begin{aligned} Cov(\mathbf{y}, \mathbf{r_e}) &= Cov(\widehat{\mathbf{y}} + \mathbf{e}, \mathbf{r_e}) \\ &= Cov(\alpha + \widehat{\beta}_G \mathbf{x} + \mathbf{e}, \mathbf{r_e}) \\ &= \widehat{\beta}_G Cov(\mathbf{x}, \mathbf{r_e}) + Cov(\mathbf{e}, \mathbf{r_e}) \\ &= Cov(\mathbf{e}, \mathbf{r_e}) \end{aligned} \tag{1.24}$$

Dans le cadre d'un modèle de régression généralisé,

$$\mathbf{y} = \boldsymbol{X}\ \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

l'approche paramétrique de la régression Gini consiste à déterminer le vecteur suivant :

$$\hat{\boldsymbol{\beta}}_G = \arg\min \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} |e_i - e_j| \right\} = \arg\min \{Cov(\mathbf{e}, \mathbf{r_e})\} \tag{1.25}$$

Il n'existe pas dans ce cas de formes fonctionnelles fermées pour l'estimateur $\hat{\boldsymbol{\beta}}_G$, la méthode est simplement numérique.

La régression Gini non paramétrique

La méthode des moindres carrés ordinaires reste l'une des méthodes les plus utilisées pour estimer la relation entre une ou plusieurs variables explicatives et une variable dépendante. Soit le modèle linéaire simple $\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x} + \boldsymbol{\epsilon}$, le coefficient de la pente de la droite de régression des moindres carrés ordinaires peut être obtenu par :

$$\begin{aligned} Cov(\mathbf{y}, \mathbf{x}) &= Cov(\beta\ \mathbf{x} + \boldsymbol{\epsilon}, \mathbf{x}) \\ &= Cov(\beta\ \mathbf{x}, \mathbf{x}) + Cov(\boldsymbol{\epsilon}, \mathbf{x}) \\ &= Cov(\beta\ \mathbf{x}, \mathbf{x}) \\ Cov(\mathbf{y}, \mathbf{x}) &= \beta\ Cov(\ \mathbf{x}, \mathbf{x}) \qquad (\mathbf{H5}) \end{aligned}$$

D'où :

$$\beta = \frac{Cov(y, \mathbf{x})}{Cov(\mathbf{x}, \mathbf{x})}$$

La régression Gini non paramétrique est construite en remplaçant la covariance usuelle par l'opérateur de Gini covariance :

$$\beta_G = \frac{Cov(\mathbf{y}, \mathbf{F}(\mathbf{x}))}{Cov(\mathbf{x}, \mathbf{F}(\mathbf{x}))} \tag{1.26}$$

Un estimateur du coefficient de la pente est obtenu en utilisant le vecteur rang mesuré au point moyen en cas d'ex aequo :

$$\hat{\beta}_G = \frac{Cov(\mathbf{y}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \tag{1.27}$$

Il est possible de montrer que l'estimateur est sans biais en supposant que $Cov(\boldsymbol{\epsilon}, \mathbf{r_x}) = 0$.
En remplaçant dans la formule de $\widehat{\beta}_G$, $\mathbf{y}$ par $\beta_G \mathbf{x} + \boldsymbol{\epsilon}$, on obtient :

$$\begin{aligned}
\widehat{\beta}_G &= \frac{Cov(\beta_G \mathbf{x} + \boldsymbol{\epsilon}\ ,\ \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\
\widehat{\beta}_G &= \frac{Cov(\beta_G\ \mathbf{x}\ ,\ \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} + \frac{Cov(\boldsymbol{\epsilon}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\
\widehat{\beta}_G &= \beta_G \frac{Cov(\mathbf{x}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} + \frac{Cov(\boldsymbol{\epsilon}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\
\widehat{\beta}_G &= \beta_G \quad \text{car } Cov(\boldsymbol{\epsilon}, \mathbf{r_x}) = 0
\end{aligned} \tag{1.28}$$

L'approche par l'opérateur de Gini covariance est considérée comme non paramétrique puisque la méthode ne nécessite pas de spécifier un modèle, ne nécessite aucune hypothèse sur les distributions, et ne nécessite aucune méthode d'optimsation. Elle est néanmoins similaire dans sa structure à la méthode des moindres carrés ordinaires.

La méthode non paramétrique (Gini et MCO) peut être vue comme une méthode géométrique où les estimateurs sont des moyennes pondérés des pentes de la courbe de régression entre chaque paire d'observations (y_i, x_i). La différence entre la régression Gini et celle des MCO se trouve dans la structure des poids attachés aux coefficients des pentes.

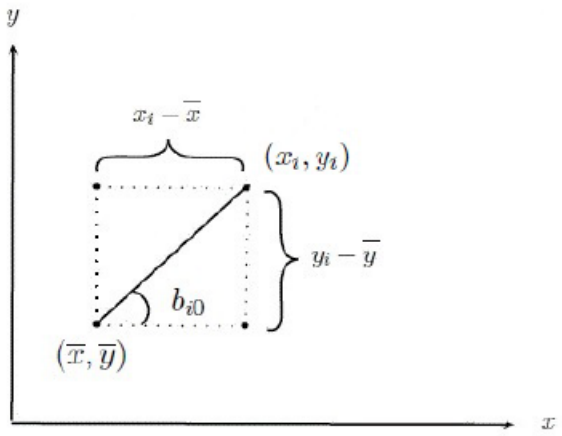


FIGURE 1.1 – Tangentes

Soit (x_i, y_i) tel que $i = 1, \ldots, n$, un échantillon provenant d'une distribution bivariée de moments d'ordre 1 et 2 finis, tel que $x_1 \leq x_2 \leq \cdots \leq x_n$. Les tangentes formées par les couples d'observations (i, j) sont :

$$m_{ij} = \frac{y_i - y_j}{x_i - x_j}, \tag{1.29}$$

Construisons maintenant un estimateur du coefficient de pente du modèle de regression par une moyenne pondérée :

$$\hat{\beta}* = \sum_{i,j} p_{ij}\, m_{ij}, \quad \text{avec} \quad \sum_{i,j} p_{ij} = 1 \tag{1.30}$$

Le schéma (ou choix) de pondération (p_{ij}) déterminera les propriétés de l'estimateur. L'estimateur par MCO se définit comme suit :

$$\widehat{\beta} = \sum_{i>j} \underbrace{\frac{(x_i - x_j)^2}{\sum_{i>j}(x_i - x_j)^2}}_{p_{ij}} \underbrace{\frac{(y_i - y_j)}{(x_i - x_j)}}_{m_{ij}} \tag{1.31}$$

Avec une pondération quadratique, les observations extrêmes (très éloignées de la moyenne) auront une influence non négligeable même pour des échantillons de grandes tailles. En effet, elles verront leur distance s'amplifier et elles auront donc un poids beaucoup plus important. La présence de

ce point prépondérant dans le cas des MCO peut donner une autre allure à la droite de regression et par conséquent fausser les interprétations.

Il serait donc judicieux de trouver un estimateur avec une pondération moins sensible aux valeurs extrêmes. L'estimateur de la pente du modèle de régression par la méthode de Gini $\widehat{\beta}_G$ peut s'écrire aussi sous forme de pondération, comme le proposent Olkin et Yitzhaki (1992). Néanmoins certains auteurs avaient déjà privilégié la piste des estimateurs obtenus par moyenne pondérée. L'estimateur proposé par Scholz (1977) et Sievers (1978) est défini comme la médiane pondérée des tangentes. Comme le montrent Olkin et Yitzhaki (1992, théorème 1), il a les mêmes propriétés que celles de la régression Gini par minimisation. En utilisant les moyennes pondérées des tangentes à la place des médianes, les auteurs définissent l'estimateur de la régression Gini non paramétrique :

$$\widehat{\beta}_G = \sum_{i>j} \underbrace{\frac{(x_i - x_j)}{\sum_{i>j}(x_i - x_j)}}_{p_{ij}} \underbrace{\frac{(y_i - y_j)}{(x_i - x_j)}}_{m_{ij}} \tag{1.32}$$

La pondération m_{ij} confère à l'esimateur $\widehat{\beta}_G$ une propriété remarquable, sur laquelle nous reviendrons, il est moins sensible aux valeurs extrêmes que l'estimateur par MCO. Dans le cas d'une régression multiple l'estimateur est donné par :

$$\widehat{\boldsymbol{\beta}}_G = (\mathbf{r_X}^T \boldsymbol{X})^{-1} (\mathbf{r_X}^T y)$$

avec $\mathbf{r_X}$ la matrice rang comportant en colonne les vecteurs rang des régresseus. Si les vecteurs rangs sont non corrélés au terme d'erreur, alors ils peuvent être utilisés comme instruments et $\mathbf{r_X}$ correspond à la matrice $\boldsymbol{Z}$ d'une régression sur variables instrumentales. La régression Gini peut donc être vue comme une régression par variables instrumentales :

$$\begin{aligned} \widehat{\beta}_{VI-MCO} &= \widehat{\beta}_G \\ (\boldsymbol{Z}^T \boldsymbol{X})^{-1} (\boldsymbol{Z}^T y) &= (\mathbf{r_X}^T \boldsymbol{X})^{-1} (\mathbf{r_X}^T y) \end{aligned} \tag{1.33}$$

Néanmoins, l'estimateur $\widehat{\beta}_G$ ne nécessite aucune hypthèse particulière contrairement à celui des MCO qui repose sur les hypothèses **H1-H5**.

Outliers

La présence d'outliers dans les données nécessite un traitement approprié pour éviter de produire des résultats peu fiables (estimateurs biaisés

et/ou non convergents).

Toute valeur qui présente des aberrations, valeur trop élevée ou trop faible, et qui ne passe pas inaperçue (pour des échantillons de faible taille) est considérée comme valeur aberrante. Les valeurs aberrantes peuvent être dues par exemple à une mauvaise saisie, ce que l'on appelle communément erreurs de mesure. A ne pas confondre avec les outliers ou valeurs extrêmes qui peuvent être considérés comme valeurs aberrantes mais avec la particularité d'être des valeurs exactes qui ne doivent pas être retirées de la base de données. Il s'agit ici d'individus qui peuvent présenter des caractéristiques atypiques sur l'une des variables étudiées. La présence de valeurs extrêmes dans la base de données peut changer l'amplitude et les signes des coefficients [1], aboutir à de mauvaises interprétations et demande l'utilisation d'une technique statistique appropriée afin d'obtenir une modélisation économétrique pertinente. En présence d'outliers, le critère des MCO ne procure plus d'estimateur stable puisque l'écart-type estimé de l'erreur tend vers l'infini ($\widehat{\sigma}_\epsilon \longrightarrow \infty$) du fait de la violation de l'hypothèse **H3**. La variance estimée de l'estimateur $\widehat{V}(\widehat{\beta})$ tend à croître de manière disproportionnée si bien que l'estimateur devient peu fiable et s'accompagne d'une statistique de $Student$ qui tend vers zéro, augmentant ainsi la probabilité d'accepter à tort l'hypothèse nulle.

Différentes méthodes de détection des outliers existent comme les méthodes statistiques inférentielles qui consistent à créer un intervalle de confiance $[\pm\, x\, ecarts{-}types]$ autour de la moyenne. Sera considérée comme valeur aberrante ou valeurs extrême, toute valeur qui se trouve hors de cet intervalle. Il est couramment utilisé dans les méthodes de détection non-automatiques des représentations graphiques des données :
1- La boîte à moustache : la détection varie selon l'amplitude x associée à l'écart-type ou à l'étendue interquartile [2] du bord de la boîte.
2- Le nuage de points : sont soupçonnés d'être des valeurs extrêmes, les points éloignés des autres points.
Il est plus prudent de ne pas se fier uniquement à ces représentations graphiques. Il est nécessaire de confirmer nos soupçons par des tests statistiques. L'un des tests les plus connus est celui de Grubbs (Grubbs, 1969 et Stefansky, 1972). Il est est utilisé pour détecter un outlier dans une distribution univariée qui suit approximativement une distribution normale :

1. Knorr et Ng(1998), Ramsawmy et al. (2000), Choi (2009).
2. La différence entre le troisième et le premier quartile (Q3 - Q1).

$$\left\| \begin{array}{l} H_0 : \text{ Présence d'outliers} \\ H_1 : \text{Absence d'outliers.} \end{array} \right.$$

La statistique de test de Grubbs est :

$$Gb = \frac{max|y_i - \bar{y}|}{s_y},$$

Le test statistique de Grubbs est le plus grand écart absolu à la moyenne de l'échantillon par unité d'écart-type. L'hypothèse d'absence d'outliers dans les données est rejetée si :

$$Gb > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/2}}{n-2+t^2_{\alpha/2}}}$$

où $t_{\alpha/2}$ représente la valeur critique de la distribution de Student à $n-2$ degrés de liberté.

La méthode du point de levier quant à elle se base aussi sur une fonction distance. Une observation qui a une valeur extrême sur une variable prédictive est appelée un point avec un effet de levier élevé. Dans un modèle de régression linéaire, le score de levier pour la $i$ème unité de données se définit comme : $h_i = \mathbf{x}_i \, (\boldsymbol{X}'\boldsymbol{X})^{-1} \, \mathbf{x}_i'$. Les éléments diagonaux h_i de la matrice $H = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ sont appelés les leviers et déterminent l'influence de la $i$ème observation sur les estimations obtenues par régression. Les valeurs de levier sont comprises entre 0 et 1, et la somme des h_i est égale au nombre de paramètres estimés. L'observation qui a une valeur levier $h_i > 2K/n$ est considérée comme aberrante.

1.4 Erreurs de mesure sur les variables

Les variables explicatives sont supposées être observées sans erreurs de mesure, ce qui s'apparente au respect de l'hypothèse **H5**. Il peut arriver dans la pratique que cette hypothèse ne soit pas vérifiée lorsque des valeurs aberrantes contaminent l'échantillon. Dans cette section, nous examinons les cas où la variable expliquée et les variables explicatives sont entachées d'erreurs de mesure.Nous dévelopons tout d'abord l'erreur de mesure sur la variable à expliquer, ensuite sur la variable explicative, et enfin le traitement par variables instrumentales.

Erreurs sur la variable endogène

Soit le modèle linéaire centré suivant $\mathbf{y} = \beta\ \mathbf{x} + \boldsymbol{\epsilon}$ tel que :

$$\begin{aligned} y_i &= \beta\, x_i + \epsilon_i \quad \text{(la } i \text{ ème observation de } y \text{ sans erreur)} \\ y_i^* &= y_i + \theta_i \quad \text{(la } i \text{ ème observation de } y \text{ avec erreur)} \\ y_i^* &= \beta\, x_i + (\epsilon_i + \theta_i) \quad (Cov(x_i, \epsilon_i) = 0) \end{aligned} \tag{1.34}$$

L'estimateur $\widehat{\beta}$ de β est par conséquent égal à :

$$\begin{aligned} \widehat{\beta} &= \frac{\sum x_i\, y_i^*}{\sum x_i^2} \\ &= \frac{\sum x_i\, [\, \beta\, x_i + (\epsilon_i + \theta_i)\,]}{\sum x_i^2} \\ &= \beta\, \frac{\sum x_i^2}{\sum x_i^2} + \frac{\sum x_i\, \epsilon_i}{\sum x_i^2} + \frac{\sum x_i\, \theta_i}{\sum x_i^2} \\ \widehat{\beta} &= \beta + \frac{\sum x_i\, \theta_i}{\sum x_i^2} \end{aligned} \tag{1.35}$$

Si $Cov(x_i, \theta_i) = 0$, alors l'estimateur $\hat{\beta}$ est sans biais puisque que par hypothèse $Cov(x_i, \epsilon_i) = 0$. De même il est convergent, mais il existe une perte d'efficacité car la variance de l'erreur est plus forte comparé au cas où la variable y_i est observée sans erreur. De même la variance de l'estimateur $\widehat{\beta}$ est plus forte lorsque y_i comprend des erreurs de mesures.

Dans la littérature, en présence d'outliers ou d'erreurs d'observations dans la variable dépendante, une méthode préconisée [3] est le *Least Absolute Deviations* (LAD). Pour mettre en évidence la pertinence de l'utilisation du LAD, nous proposons de simples simulations de Monte Carlo où nous déduisons l'erreur quadratique moyenne (MSE) des coefficients estimés $\widehat{\beta}$ lorsque la contamination de y porte sur une seule observation.

3. A titre d'exemple, voir Dodge (1997).

Algorithm 1: Simulations de Monte Carlo

Result: Méthode robuste du LAD en présence d'outliers dans y

Générer des variables $\boldsymbol{X} \sim \mathcal{N},\ \boldsymbol{\varepsilon} \sim \mathcal{N}$;

Déduire la variable $y = \mathbf{1}\alpha + \beta\mathbf{x} + \boldsymbol{\varepsilon}$, en fixant $\beta = 10$;

$\theta = 50$ [θ est la valeur de l'erreur d'observation], $n = 1000$ et ;

$i = 1$ [i est le nombre d'itérations];

repeat

  Déduire la variable y_l en introduisant l'outlier uniquement dans la ligne l de y : $y_l^* = y_l + \theta$;

  Calculer et récupérer le coefficient $\widehat{\beta}$ issu de de la regression y/x par les méthodes MCO, Gini et LAD;

until $i = 1000$ [par pas de 1];

return Mean squared Errors (MSE) du coefficient β selon les trois méthodes ;

Les résultats sont les suivants.

MSE MCO	MSE Gini	MSE LAD
79640.81	78515.66	48054.62

TABLE 1.1 – Comparaison des Mean Squared Error (MSE)

Ce tableau confirme à travers le calcul des MSE que la méthode LAD se démarque des autres méthodes lorsqu'il existe des erreurs de mesure au niveau de la variable dépendante. Pour illustrer ces résultats, prenons un échantillon de dix observations et introduisons arbitrairement à la dixième observation un outlier dans de y.

Ce graphique illustre la robustesse de l'estimateur LAD mais aussi le fort impact que la présence d'un outlier dans la variable y peut engendrer sur l'estimateur de la regression Gini. En effet soit les modèles centrés sans erreur et contaminé suivants où θ est un vecteur de même taille que y :

$$\begin{aligned} y &= \beta_G\, \mathbf{x} + \boldsymbol{\varepsilon} \\ y^* &= y + \theta \end{aligned}$$

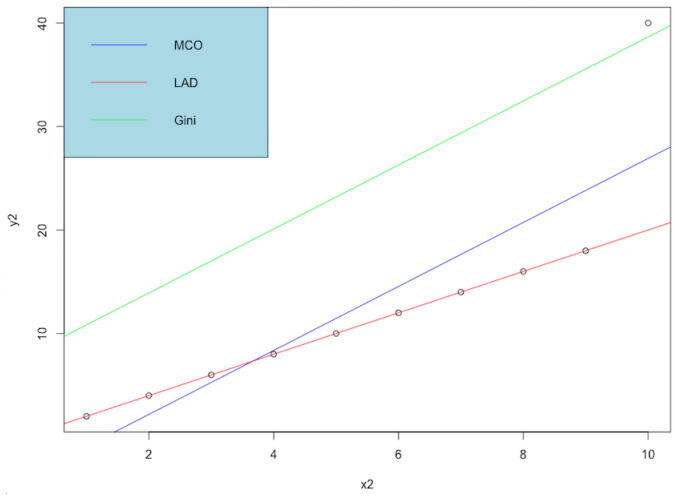


FIGURE 1.2 – Illustration graphique de la présence d'outlier dans y

L'estimateur non paramétrique de la régression Gini contaminée est :

$$\begin{aligned}
\widehat{\beta^{\theta}}_G &= \frac{Cov(y\ +\ \theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\
\widehat{\beta^{\theta}}_G &= \frac{Cov(y, \mathbf{r_x}) + Cov(\theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\
\widehat{\beta^{\theta}}_G &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} + \frac{Cov(\theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\
\widehat{\beta^{\theta}}_G &= \widehat{\beta}_G + \frac{Cov(\theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})}
\end{aligned}$$

L'estimateur est sans biais si, et seulement si, $Cov(\theta, \mathbf{r_x}) = 0$. Lorsque $\widehat{\beta}_G$ est positif il peut donc être biaisé par le bas ou par le haut :

$$\begin{cases} \widehat{\beta^{\theta}}_G > \widehat{\beta}_G & \text{ssi } Cov(\theta, \mathbf{r_x}) > 0 \\ \widehat{\beta^{\theta}}_G < \widehat{\beta}_G & \text{ssi } Cov(\theta, \mathbf{r_x}) < 0 \end{cases}$$

Erreurs sur les variables explicatives

Soit le modèle avec les erreurs suivantes :

$$\begin{aligned}
y_i &= \beta\, x_i + \epsilon_i \quad \text{(avec } x_i \text{ la } i \text{ ème valeur de x correctement observée)} \\
x_i^* &= x_i + \theta_i \quad \text{(la } i \text{ ème observation contaminée de x)} \\
y_i &= \beta\, x_i^* + \epsilon_i \quad \text{avec } \mathbb{E}(x_i^*\, \epsilon_i) = 0 \\
y_i &= \beta\, (x_i + \theta_i) + \epsilon_i \\
y_i &= \beta\, x_i + \epsilon_i + \beta\, \theta_i \\
y_i &= \beta\, x_i + \epsilon_i^* \text{ avec } \epsilon_i^* = (\epsilon_i + \beta\, \theta_i)
\end{aligned} \tag{1.36}$$

Nous faisons les hypothèses usuelles suivantes $\theta_i \rightsquigarrow N(0, \sigma_\theta)$, $\mathbb{E}(\theta_i\, x_i^*) = 0$, et $\mathbb{E}(\theta_i\, \epsilon_i) = 0$. Dans l'expression de y_i (ligne 3), nous avons : $\mathbb{E}(x_i^*\, \epsilon_i) = 0$ mais $\mathbb{E}(x_i\, \epsilon_i^*) \neq 0$.

Preuve

$$\begin{aligned}
\mathbb{E}(x_i\, \epsilon_i^*) &= \mathbb{E}(\, (x_i^* - \theta_i)(\epsilon_i + \beta\, \theta_i)\,) \\
&= \mathbb{E}(\, x_i^* \epsilon_i + \beta \theta_i x_i^* - \theta_i \epsilon_i - \beta\, \theta_i^2\,) \\
&= \mathbb{E}(\, -\beta\, \theta_i^2\,) \\
&= -\beta\, \mathbb{E}(\, \theta_i^2\,) \\
\mathbb{E}(x_i\, \epsilon_i^*) &= -\beta\, \sigma_\theta^2
\end{aligned} \tag{1.37}$$

La variable explicative sans erreur de mesure est par conséquent corrélée avec le terme d'erreur, car rappelons que le modèle estimée est : $y_i = \beta\, x_i + \epsilon_i^*$. Par conséquent, nous avons les implications de la corrélation entre la variable explicative et le terme d'erreur (**H5**). L'estimateur ($\hat{\beta}$) du modèle est biaisé et non convergent. En effet, nous avons :

$$
\begin{aligned}
\widehat{\beta} &= \frac{\sum x_i\, y_i}{\sum x_i^2} \\
&= \frac{\sum x_i\, (\beta\, x_i + \epsilon_i^*)}{\sum x_i^2} \\
&= \frac{\sum \beta\, x_i^2 \;+\; x_i \epsilon_i^*}{\sum x_i^2} \\
&= \beta \;+\; \frac{\sum x_i \epsilon_i^*}{\sum x_i^2} \\
&= \beta \;+\; \frac{\sum (x_i^* - \theta_i)\, (\epsilon_i + \beta\, \theta_i)}{\sum x_i^2} \\
&= \beta \;+\; \frac{\sum (x_i^*\, \epsilon_i)}{\sum x_i^2} \;+ \beta \frac{\sum (x_i^*\, \theta_i)}{\sum x_i^2} \;-\; \frac{\sum (\theta_i\, \epsilon_i)}{\sum x_i^2} - \beta\, \frac{\sum \theta_i^2}{\sum x_i^2} \\
\widehat{\beta} &= \beta \;-\beta\, \frac{\sum \theta_i^2}{\sum x_i^2}
\end{aligned}
\tag{1.38}
$$

La limite en probabilité de $\widehat{\beta}$ est donc :

$$
\text{plim}(\hat{\beta}) = \beta \;-\; \beta\, \frac{s^2_{(\theta_i)}}{s^2_{(x_i)}} \tag{1.39}
$$

Avec respectivement $s^2_{(\theta_i)}$ et $s^2_{(x_i)}$ les variances estimées de θ_i et x_i.
L'estimateur $\hat{\beta}$ est non convergent et biaisé négativement. Montrons par simulation de Monte Carlo la robustesse de la regression Gini, lorsqu'il existe une erreur de mesure dans une seule observation de la variable explicative.

Algorithm 2: Simulations de Monte Carlo

Result: Robustesse du Gini en présence d'erreur de mesure dans $\mathbf{x}$

Générer des variables $\boldsymbol{X} \sim \mathcal{N}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}$;

Déduire la variable $y = \alpha + \beta\mathbf{x} + \boldsymbol{\varepsilon}$, en fixant $\beta = 10$;

$\theta = 50$ [θ est la valeur de l'erreur d'observation], $n = 1000$ et ;

$i = 1$ [i est le nombre d'itérations];

repeat

  Générer une variable $\mathbf{x}$ en introduisant l'outlier uniquement dans la ligne l de $\mathbf{x}$: $x_l = x_l^* + \theta$;

  Calculer le vecteur rang de $\mathbf{x}$;

  Calculer et récupérer le coefficient $\widehat{\beta}$ issu de de la regression $y/\mathbf{x}$ par les méthodes des MCO, Gini et LAD;

until $i = 1000$ [par pas de 1];

return Mean squared Errors (MSE) du coefficient β selon les trois méthodes ;

Les résultats sont les suivants.

MSE MCO	MSE Gini	MSE LAD
66.73	12.11	39.24

TABLE 1.2 – Comparaison des Mean Squared Error (MSE)

Le tableau montre que l'estimateur de la régression Gini non paramétrique a le plus faible MSE et donc semble être une méthode appropriée pour les erreurs de mesures en $\mathbf{x}$. Pour illustrer ce résultat, générons une droite de régression avec erreur de mesure à la dixième et dernière observation.

Ce graphique illustre la robustesse de la méthode Gini grâce à l'utilisation du vecteur rang. Supposons que ce dernier, $\mathbf{r_x}$, reste inchangé après l'introduction d'un vecteur d'outliers θ qui contamine l'ensemble des valeurs de $\mathbf{x}$ tel que $\mathbf{x}^* = \mathbf{x} + \theta$, alors l'estimateur contaminé de β_G est :

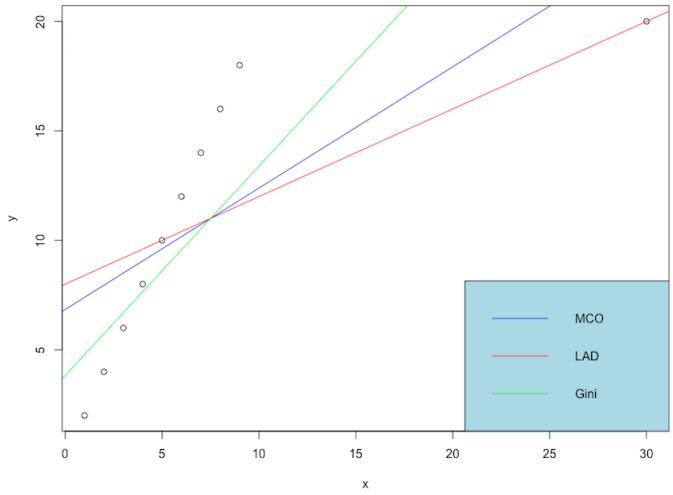


FIGURE 1.3 – Illustration graphique de la présence d'outlier dans x

$$\begin{aligned}
\widehat{\beta^{\theta}}_{G} &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}^*, \mathbf{r_x})} \\
\widehat{\beta^{\theta}}_{G} &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x} + \theta, \mathbf{r_x})} \\
\widehat{\beta^{\theta}}_{G} &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x}) + Cov(\theta, \mathbf{r_x})}
\end{aligned}$$

Puisque,

$$\begin{aligned}
\text{plim} \left[\frac{1}{n} y' \mathbf{r_x} \right] &= Cov(y, \mathbf{r_x}) \\
\text{plim} \left[\frac{1}{n} \mathbf{x}' \mathbf{r_x} \right] &= Cov(\mathbf{x}, \mathbf{r_x}) \\
\text{plim} \left[\frac{1}{n} \theta' \mathbf{r_x} \right] &= Cov(\theta, \mathbf{r_x})
\end{aligned}$$

alors :

$$\text{plim}\, \widehat{\beta^{\theta}}_{G} = \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x}) + Cov(\theta, \mathbf{r_x})}$$

Par conséquent :
Si $Cov(\theta, \mathbf{r_x}) > 0 \Rightarrow \widehat{\beta^{\theta}}_G < \widehat{\beta}_G$.
Si $Cov(\theta, \mathbf{r_x}) < 0 \Rightarrow \widehat{\beta^{\theta}}_G > \widehat{\beta}_G$.

L'erreur d'observation dans la variable explicative x entraîne une contamination de l'estimateur du Gini caractérisé par le terme $Cov(\theta, \mathbf{r_x})$. Dans le cadre de la regression par MCO, la contamination provient du terme $Cov(\theta, \mathbf{x})$. Par conséquent, si $Cov(\theta, \mathbf{x}) > Cov(\theta, \mathbf{r_x})$, l'impact de l'erreur de mesure dans x est plus faible avec la regression Gini.
Pour identifier ce problème, un test d'exogénéité de type Hausman (1978) peut permettre de détecter un terme d'erreur attaché à une (ou plusieurs) variables explicatives. En cas de violation de l'hypothèse **H5**, le recourt à la méthode par variables instrumentales s'impose.

La méthode des variables instrumentales

Deux approches sont développées dans cette scetion, celle de la méthode des variables instrumentales par MCO (VI-MCO) et celle des variables instrumentales avec le rang des instruments (Gin-VI). L'avantage d'utiliser le rang est la possibilité d'obtenir de plus amples informations sur la relation qui lie les données, comme celle de savoir si la relation entre la variable dépendante et les variables explicatives ou celle entre les variables explicatives elles-mêmes est monotone ou non. Soit le modèle général :

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

tel que plim$[\mathbf{X}'\,\boldsymbol{\epsilon}/N] \neq 0$. La technique des variables instrumentales par MCO bien connue consiste à trouver des instruments à la fois corrélées aux variables explicatives et non corrélées au terme d'erreur. Nous considérons dans ce qui suit que le nombre de variables instrumentales est égal à celui des variables explicatives. Les principes de base de cette régression sont les suivants :
1) Asymptotiquement, il n'y a pas de corrélation entre les instruments et le terme d'erreur : pLim$(\mathbf{Z}'\,\boldsymbol{\epsilon}/n) = 0$
2) Il existe une forte corrélation entre les variables instrumentales $\boldsymbol{Z}$ et les variables explicatives $\boldsymbol{X}$: plim$(\boldsymbol{Z}'\boldsymbol{X}/n) = \tau$.

3) Les variables instrumentales $\boldsymbol{Z}$ ne sont pas colinéaires : $\text{plim}(\boldsymbol{Z}'\boldsymbol{Z}/n) = \tau^*$.

Rappelons que l'on peut déduire les estimateurs VI-MCO de deux manières différentes : une application directe d'une matrice d'instruments ; ou l'utilisation des moindres carrés en deux étapes. Pour la méthodologie Gini-VI, les deux méthodes peuvent donner des estimateurs totalement différents. Nous présentons dans un premier temps, les deux méthodes relatives aux MCO puis celles relatives au Gini.

L'estimateur VI-MCO est celui de l'estimateur MCO du modèle transformé $\boldsymbol{Z}'\mathbf{y} = \boldsymbol{Z}'\boldsymbol{X}\ \boldsymbol{\beta} + \boldsymbol{Z}'\boldsymbol{\varepsilon}$:

$$\hat{\beta}_{VI-MCO1} = (\boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\mathbf{y}$$

or $\mathbf{y} = \boldsymbol{X}\ \boldsymbol{\beta} + \boldsymbol{\epsilon}$, alors :

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= (\boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'(\boldsymbol{X}\ \boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= (\boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{X})^{-1}\ (\boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{X})\ \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{\epsilon} \\
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{\epsilon}
\end{aligned}$$

La limite en probabilité de l'estimateur $\hat{\boldsymbol{\beta}}_{VI-MCO1}$ est $\boldsymbol{\beta}$ car $\text{plim}(\boldsymbol{Z}'\boldsymbol{\epsilon}/n) = 0$. L'estimateur est sans biais. L'expression simplifiée de l'estimateur VI-MCO est, lorsque le nombre d'instruments est égal au nombre de variables explicatives ($\boldsymbol{Z}'\boldsymbol{X}$ est une matrice carrée) :

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= (\boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\boldsymbol{Z}\ \boldsymbol{Z}'\mathbf{y} \\
&= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\ \overbrace{(\boldsymbol{X}'\boldsymbol{Z})^{-1}\ \boldsymbol{X}'\boldsymbol{Z}}^{=Id}\ \boldsymbol{Z}'\mathbf{y} \\
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\ \boldsymbol{Z}'\mathbf{y}
\end{aligned} \tag{1.40}$$

L'estimateur VI en deux étapes consiste d'abord à faire un modèle de type $\boldsymbol{X} = \boldsymbol{Z}\pi + \mathbf{r}$ où on estime π afin d'estimer $\hat{\boldsymbol{X}}$ que l'on utilise dans la deuxième étape qui consiste à régresser $\mathbf{y}$ sur $\widehat{\boldsymbol{X}}$. On retrouve ainsi $\hat{\beta}_{VI-MCO1} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\ \boldsymbol{Z}'\mathbf{y}$.

En 2004, Yitzhaki et Schechtman proposent une régression Gini avec variables instrumentales communément appelée Gini-VI. Comme dans le cas des MCO, deux méthodes permettent d'obtenir deux estimateurs, mais

nous verrons qu'ils ne sont pas nécessairement identiques. La méthode directe Gini-VI est :

$$\hat{\beta}_{Gini-VI1} = (\boldsymbol{R}'_{\boldsymbol{Z}}\ \boldsymbol{X})^{-1}\ \boldsymbol{R}'_{\boldsymbol{Z}}\ \mathbf{y} \tag{1.41}$$

oú $\boldsymbol{R_Z}$ est la matrice des rangs des variables instrumentales. La procédure en deux étapes se fait de la manière suivante. Premièrement :

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{R}'_{\boldsymbol{Z}})^{-1}\boldsymbol{R}'_{\boldsymbol{Z}}\boldsymbol{X} \tag{1.42}$$

Dans la seconde étape :

$$\begin{aligned}
\hat{\beta}_{Gini-VI2} &= (\boldsymbol{R}'_{\hat{\boldsymbol{X}}}\ \hat{\boldsymbol{X}})^{-1}\ \boldsymbol{R}'_{\hat{\boldsymbol{X}}}\mathbf{y} \\
&= [\ (\boldsymbol{R}'_{\hat{\boldsymbol{X}}}\boldsymbol{Z})\ (\boldsymbol{R}'_{\boldsymbol{Z}}\boldsymbol{Z})^{-1}\ \boldsymbol{R}'_{\boldsymbol{Z}}\boldsymbol{X}\ ]^{-1}\ \boldsymbol{R}'_{\hat{X}}\mathbf{y} \\
\hat{\beta}_{Gini-VI2} &= (\boldsymbol{R}'_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\ \boldsymbol{R}'_{\boldsymbol{Z}}\boldsymbol{Z}\ (\boldsymbol{R}'_{\hat{\boldsymbol{X}}}\boldsymbol{Z})^{-1}\ \boldsymbol{R}'_{\hat{\boldsymbol{X}}}\mathbf{y}
\end{aligned}$$

Nous constatons que dans la méthode directe, nous avons la matrice $\boldsymbol{R}$ qui est la matrice des rangs des variables instrumentales ($\boldsymbol{R_Z}$), tandis que dans la méthode en deux étapes, deux matrices $\boldsymbol{R}$ sont utilisées, celle des rangs de $\hat{\boldsymbol{X}}$ et celle de $\boldsymbol{Z}$. Par conséquent, à moins que les rangs de tous les instruments et de la variable explicative initiale soient identiques, nous devrions nous attendre à des résultats différents. Cela peut s'expliquer par le fait que la fonction de répartition n'est pas en général une transformation linéaire de la variable explicative. Et donc, tout se passe comme si nous espérions avoir une relation linéaire malgré que nous utilisons une transformation non linéaire.

Il est comparé dans l'article de Yitzhaki et Schechtman (2004), les propriétés de la méthode d'estimation des variables instrumentales sous deux paramètres alternatifs (VI-MCO et Gini-VI). Le premier paramètre, dit paramètre standard (VI-MCO) est basé sur la minimisation d'une fonction quadratique de l'erreur, comme dans le cas de la régression par les MCO. Le second estimateur (Gini-VI) quant à lui, est une méthode d'estimation des variables instrumentales par le bias du GMD comme méthode de régression.
Ainsi les coefficients de régression de ces différentes approches de la méthode des variables instrumentales peuvent être interprétés comme des moyennes pondérées des pentes entre les observations adjacentes. Par conséquent, face aux limites de la variance et de la régression par les MCO,

il apparait que l'estimateur de la méthode des variables instrumentales issu de la régression Gini est moins sensible aux outliers et à la violation de l'hypothèse de linéarité que l'estimateur standard de la méthode des variables instrumentales.
Il est à noter que si le modèle utilisé est linéaire et que les hypothèses communément utilisées sont respectées, alors les deux méthodes produisent les mêmes estimateurs.

1.5 Inférence statistique

Le problème posé est de savoir quelle est la loi suivie par les estimateurs issus de la regression Gini $\widehat{\beta}_G$ et de déterminer les propriétés de ces derniers, notamment la convergence. Yitzhaki et Schechtman (2013) montrent que les estimateurs issus de la régression Gini sont des U-statistiques. La théorie des U-statistiques est une théorie d'échantillonnage qui permet de déterminer, sous des conditions peu exigeantes, que l'estimateur est sans biais et convergent (Hoeffding, 1948). Cette théorie permet par ailleurs de déterminer la variance des estimateurs et d'effectuer des tests statistiques de signification asymptotiques sur ces derniers.

Si la statistique dont on cherche l'estimateur s'écrit :

$$\theta(F) = \mathbb{E}[\phi(X_1, \cdots, X_m)] = \int \dots \int \phi(x_1, \dots, x_m) dF(x_1) \cdots dF(x_m),$$

où $\phi(x_1, \dots, x_m)$ est appelé le noyau de $\theta(F)$ de degré m. Alors l'estimateur U (sans biais) de θ existe tel que $U \sim \mathcal{N}$:

$$U_n := \binom{n}{m}^{-1} \sum_c \phi(X_{i_1}, X_{i_2}, \dots, X_{i_m})$$

où $\sum_c$ indique la somme pour toutes les combinaisons de m éléments $\{i_1, \cdots, i_m\}$ de $\{1, \cdots, n\}$.

Les conditions qui garantissent la convergence de U et l'absence de biais sont les suivantes :

- $\phi(x_1, \dots, x_m)$ doit être une fonction symétrique ;
- les variables aléatoires X_i doivent être $i.i.d.$;
- la taille de l'échantillon doit être suffisamment large.

Prenons quelques exemples usuels.

Exemple 1.5.1. *Soit $\theta(F_X) = \mathbb{E}(X)$. Le noyau est $\phi(X) = X$, il est symétrique de degré $m = 1$. Alors la U-statistique est :*

$$U_1 = \frac{1}{\binom{n}{1}} \sum_i x_i = \frac{1}{\frac{n!}{1!\,(n-1)!}} \sum_i x_i = \frac{1}{\frac{n\,(n-1)!}{1\,(n-1)!}} \sum_i x_i = \frac{1}{n} \sum_i x_i = \bar{x}$$

Exemple 1.5.2. *Soit $\theta(F_X) = \mathbb{E}(X_1 - X_2)^2$. Le noyau est $\phi(X) = \frac{1}{2}(X_1 - X_2)^2$, il est symétrique de degré $m = 2$. La U-statistique est donc :*

$$\begin{aligned} U &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(x_i - x_j)^2}{2} \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

Revenons à la régression Gini linéaire simple. Nous avons vu que :

$$\begin{aligned} \beta_G &= \frac{Cov(\ y, F_{\mathbf{x}}\ )}{Cov(\ \mathbf{x}, F_{\mathbf{x}}\ )} \\ &= \frac{GCov(\ y, \mathbf{x}\ )}{GCov(\ \mathbf{x}, \mathbf{x}\ )} \end{aligned}$$

Proposition 1.5.1. – Yitzhaki et Schechtman (2013, Chapitre 9) – *: Soient (X_1, Y_1) et (X_2, Y_2) de dimension 2 d'une distribution bi-variée continue dont les deux premiers moments sont finis. Soit $h((X_1, Y_1), (X_2, Y_2)) = (X_1, Y_1)\mathbb{1}_{(Y_1 > Y_2)} + (X_2, Y_2)\mathbb{1}_{(Y_1 > Y_2)})$ où $\mathbb{1}_{a>b}$ est défini comme :*

$$\mathbb{1}_{a>b} = \begin{cases} 1 & \text{si } a > b \\ 0 & \text{sinon.} \end{cases}$$

Alors $h((X_1, Y_1), (X_2, Y_2))$ est un noyau symétrique de degré (2,2) pour $\Delta_{X,Y} = 4Cov(X, F(Y)) = 4GCov(X, Y)$. [4]

Preuve:
Voir Yitzhaki et Schechtman (2013, Chapitre 9, p. 203-204). ■

4. Un degré (2,2) signifie qu'on a besoin de deux X indépendants et deux Y indépendants afin d'obtenir un estimateur non biaisé.

En utilisant le noyau ci-dessus, la U-statistique est

$$\begin{aligned} U(\Delta_{X,Y}) &= \frac{1}{\binom{n}{2}} \sum_{i<j} \sum h((x_i, y_i), (x_j, y_j)) \\ &= \frac{1}{\binom{n}{2}} \sum_{i<j} \sum [(x_i - x_j)\mathbb{1}_{y_i > y_j} + (x_j - x_j)\mathbb{1}_{y_i > y_j}] \end{aligned}$$

Il s'agit d'une U-statistique pour $4Cov(X, F_Y)$ et par conséquent un estimateur sans biais et convergent. Une définition alternative, basée sur une combinaison linéaire des éléments concomitants des statistiques d'ordre, est donnée par :

$$U(\Delta_{X,Y}) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\, x_{y(i)}$$

où $x_{y(i)}$ *et* $x_{(i)}$ sont des statistiques d'ordre avec $x_{y(i)}$ la valeur de x qui correspond à la ième statistique d'ordre de $y_1, \cdots, y_n$.

Nous avons vu que :

$$\beta_G = \frac{Cov(\, y, F_{\mathbf{x}}\,)}{Cov(\, \mathbf{x}, F_{\mathbf{x}}\,)}$$

Étant donné que $Cov(\mathbf{y}, F_{\mathbf{x}})$ et $Cov(\mathbf{x}, F_{\mathbf{x}})$ sont des fonctions symétriques, alors il existe une U-statistique, telle que les estimateurs de $Cov(\mathbf{y}, F_{\mathbf{x}})$ et $Cov(\mathbf{x}, F_{\mathbf{x}})$ sont sans biais et respectivement donnés par :

$$U_1 = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\, x_{y(i)} \tag{1.43}$$

$$U_2 = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\, x_{(i)}, \tag{1.44}$$

Theorem 1.5.1. *Pour U_1 et U_2 les U-statistiques des estimateurs respectifs de la $Cov(\mathbf{y}, F_{\mathbf{x}})$ et de la $Cov(\mathbf{x}, F_{\mathbf{x}})$, l'estimateur $\widehat{\beta}_G$ du paramètre β_G est une U-statistique tel que $\widehat{\beta}_G = \frac{U_1}{U_2} \overset{a}{\sim} \mathcal{N}$.*

Preuve:
Soit :

$$\widehat{\beta}_G = \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})}$$

Des équations (43) et (44), l'estimateur $\widehat{\beta}_G$ du paramètre β_G s'écrit :

$$\widehat{\beta}_G = \frac{U_1}{U_2}$$

L'estimateur $\widehat{\beta}_G$ étant un ratio de deux U-statistiques, alors $\widehat{\beta}_G$ est aussi une U-statistique. Du théorème 10.4 de Yitzhaki et Schechtman (2013), nous avons : Soit $(U') = U_1, \cdots, U_t$ t U-statistiques basées sur un échantillon $x_1, \cdots, x_n$ de taille n avec U_i correspondant à θ_i (avec un noyau h_i), $i = 1, \cdots, t$. Si la fonction $g(y) = g(y_1, \cdots, y_t)$ qui ne comporte pas n et est continue avec ses dérivées partielles à certains voisinages du point (y) = (θ) = ($\theta_1, \cdots, \theta_t$) et sous la condition que $\mathbb{E}[h_i^2(X_1, X_2, \cdots, X_{m_i})] < \infty$ alors l'estimateur $\widehat{\beta}_G$ tend vers une loi normale lorsque $n \longrightarrow \infty$. ■

L'expression de la variance d'une U-statistique est très complexe et d'utilisation pratique assez délicate dès lors que m est supérieur à 1. Une alternative est l'utilisation de la méthode de Jackknife qui permet de réduire le biais. De manière générale, l'estimateur de la variance par Jackknife d'un estimateur U est donné par :

$$\hat{V}(\hat{\beta}_G) = \frac{n-1}{n} \sum_{i=1}^{n} \left[U_{-i} - \frac{1}{n} \sum_{i=1}^{n} U_{-i} \right]^2$$

avec U_{-i} l'estimateur d'une U-statistique sur un échantillon de taille n sans la ième observation. De cette manière, la statistique de test pour le coefficient de la pente d'une régression Gini est déduit du fait que $\hat{\beta}_G \overset{a}{\sim} \mathcal{N}(\beta_G, \widehat{V}(\hat{\beta}_G))$.

1.6 Conclusion

Ce chapitre a pour objectif l'enrichissement de la littérature sur le Gini's Mean Difference (GMD). Il met d'abord en lumière les limites de la régression des moindres carrés ordinaires (MCO) en présence d'outliers dans les observations. De même en cas d'endogénéité de l'une des variables explicatives ou d'erreur de mesure, la régression par les MCO s'avère ne pas être efficace. Ensuite, une nouvelle approche est développée, la regression Gini (Yitzhaki et Schechtman - 2013) en vue de remédier à cette sensibilité de la méthode des MCO et surtout de s'affranchir de certaines hypothèses

de base en économétrie. Cette approche avec en son coeur l'opérateur Cogini peut être perçue comme un mélange de la méthode de pur rang de Spearman (métrique ℓ_1) et la méthode de la variance (métrique ℓ_2).
En présence des points atypiques qui limitent l'utilisation des MCO, la régression Gini produit des estimateurs plus robustes que les estimateurs des MCO. De plus, les estimateurs issus de la régression Gini sont des U-statistiques qui autorisent l'inférence de différents estimateurs, permettant ainsi de tester la qualité des modèles étudiés. Les outils de cette méthodologie sont utilisés dans les chapitres suivants.

BIBLIOGRAPHIE

Alan Stuart. (1964). *Limit Distribution For Total Rank*.The British Psychological Society , **vol. 7**, Issue 1, pp. 50 - 51

Corrado Gini. (1912). *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*, C. Cuppini, Bologna.

Haim Shalit and Sholmo Yitzhaki. (1984) *Mean-Gini, portfolio Theory, and the Pricing of Risy Assets*.The Journal of Finance, vol. XXXIX, n. 5.

Hideo Tanaka et Satoru Uejima. (1982). *Linear regression analysis with fuzzy Model*.TIEEE Transaction on systems, man, and cybenetics , **vol. SMC-12 n. 6.**

Kannan Senthamarai et Kuppusamy Manoj. (2015). *Outlier Detection in Multivariate Data*. Applied Mathematical Sciences , **vol. 9**, Issue 1, pp. 50 - 51, n. 47, pp. 2317 - 2324.

Marcel Carcea and Robert Serfling. (2015). *A Gini autocovariance function for time series modeling*. Journal of Time Series Analysis, **vol. 7**, 36, 817-838.

Marie Davidian and Raymond Carroll. (1987). *Variance function estimation*. Journal of the American Statistical Association, **vol. 7**, 82, 400, (December), 1079-1091.

Ndéné Ka and Stéphane Mussard. (2015). ℓ_1 *regressions : Gini estimators for fixed effects panel data*. Journal of the American Statistical Association, **vol. 46**, n.8, pp. 1436 - 1446.

Téa Ouraga. (2018). *La Régression Gini : Une Revue de la Littérature*. Université de Nîmes, Faculté d'Économie, Nîmes, France.

Marcel Carcea and Robert Serfling. (2015). *A Gini autocovariance function for time series modeling*. Journal of Time Series Analysis, **vol. 7**, 36, 817-838.

Roger Koenker et Gilbert Bassett. (1978). *Regression quantile* . Econometrica, **vol. 46**, pp. 33 - 50.

Sholmo Yitzhaki. (2003). *Gini's Mean difference : a superior measure of variability for non-normal distributions*. METRON - International Journal of Statistics, **vol. LXI**, n. 2, pp. 285 - 316.

Shlomo Yitzhaki & Edna Schechtman. (2004). *The Gini Instrumental Variable, or the "double instrumental variable estimator* . METRON - International Journal of Statistics, **vol. LXII**, n. 3, pp. 287 - 313.

Shlomo Yitzhaki & Edna Schechtman. (2013). *The Gini Methodology : A Primer on a Statistical Methodology* . New York : Springer.

Wassily Hoeffding. (1984). *class of Statistics with Asymptotically Normal Distribution* . The Annals of Mathematical Statistics, **vol. 19**, n. 3, pp. 293 - 325.

CHAPITRE 2

A NOTE ON GINI PRINCIPAL COMPONENT ANALYSIS

Sommaire

Résumé

Une analyse en composantes principales basée sur la matrice de corrélation au sens de Gini est proposée (ACP-Gini). Il est montré que dans le cas gaussien, l'ACP classique est équivalent à l'ACP-Gini. Il est également prouvé que la réduction de dimensions basée sur la matrice de corrélation de Gini, qui repose sur les distances de Manhattan, est robuste aux valeurs aberrantes. Des simulations de Monte Carlo montrent la robustesse de l'ACP-Gini par rapport à l'ACP basée sur la variance. Ce chapitre est un article publié, *C.f* Ouraga. (2019).

Mots-clés : (R) Statistiques Multivariées ; Gini ; PCA ; Robustesse.

Abstract

A principal component analysis based on the Gini correlation matrix is proposed (PCA-Gini). It is shown that in the Gaussian case, the classical PCA is equivalent to the Gini PCA. It is also shown that the dimension reduction based on the Gini correlation matrix, which is based on city-block distances, is robust to outliers. Monte Carlo simulations show the robustness of the Gini PCA with respect to the variance-based PCA. This chapter is a published paper, *i.e* Ouraga. (2019).

Keywords : (R) Multivariate statistics ; Gini ; PCA ; Robustness.

2.1 Introduction

In 1912, Gini proposed the Gini Mean Difference index (GMD) as a new way to measure inequality and disparity between individuals in a given sample :

$$GMD_x = \mathbb{E}\left|x_i - x_j\right|, \tag{2.1}$$

where x_i and x_j are two realizations of the random variable x. The GMD is based on the taxi-cab distance and thus offers an alternative measure to the usual variance based on the euclidean metrics :

$$\sigma_x^2 = cov(x,x) = \frac{1}{2}\,\mathbb{E}\left|x_i - x_j\right|^2. \tag{2.2}$$

Since then, two main approaches have been developed in the literature for analyzing the variability between two random variables.

The first one is based on the covariance between the c.d.f. of the random variable x and that of y, expressed as :

$$S_{x,y} = cov(F(x), F(y)). \tag{2.3}$$

This is the well-known Spearman's method defined to be the rank method. The second one, the Gini approach, has been developed by Schechtman and Yitzhaki (2003) who paved the way on the covariance Gini operator – cogini operator from now on – which can be seen as a mixture of the variance and Spearman's pure rank approaches :

$$cog(x,y) = cov(x, F(y))\ ;\ cog(y,x) = cov(y, F(x)). \tag{2.4}$$

It is noteworthy that $cog(x,x) = 1/4GMD_x$, as a consequence, the cogini operator is closely related to the ℓ_1 metric.

The cogini operator has some appealing features. For instance, Olkin and Yitzhaki (1992) and Yitzhaki and Schechtman (2013) point out that the ordinary least squares method can be employed by replacing the usual covariance operator by the cogini one. Their Gini regression has been shown to be robust to outliers. Indeed, the variance criterion may be misleading to handle a sample with extreme values or to deal with heavy-tailed distributions, see Carcea and Serfling (2015) in the case of times series. Also, as shown by Greselin (2015) the use of cogini operators close to Choquet integrals may be useful to unify measures of inequality and risk.

In this chapter, we start from the recognition that the cogini lies in the family of robust statistics, and as such, it is a good candidate to perform Principal Component Analysis in the Gini sense (Gini PCA), Ouraga*i.e* (2019). In the field of PCA, Baccini *et al.* (1996) were among the first authors dealing with a ℓ_1-norm PCA framework. Their idea was to robustify the standard PCA by means of the Gini Mean Difference as an estimator of the standard deviation. Ding *et al.* (2006) made use of the R_1 norm to robustify the PCA, in which the Euclidean distance is applied over the dimensions of the matrix only, whereas Frobienus norm is concerned with the Euclidean distance applied to both dimensions and observations (rows of the matrix of the data).

The aim of this chapter is to use the cogini operator underlying the correlation Gini index in order to provide a Gini PCA less sensitive to outlying observations than the usual PCA by substituting the variance-covariance matrix to the Gini correlation matrix. Contrary to Baccini *et al.* (1996) in which the PCA is formalized by replacing the standard deviation of each variable by its GMD, we employ the Gini correlation index between all pairs of variables (Section 2.2). We show with simple Monte Carlo simulations that the Gini PCA is robust to outliers thanks to the relative and absolute contributions, that are respectively, the distance of the observations to the principal components and their contributions to the overall Gini correlation (Section 2.3). Section 3.8 closes the note.

2.2 Gini PCA

Let $\mathbf{X} = [x_{ik}]$ be a $N \times K$ matrix that describes N observations on K dimensions such that $N \gg K > 1$, with elements $x_{ik} \in \mathbb{R}$ that reports the score of observation i on dimension k. The $N \times 1$ vectors representing each column of $\mathbf{X}$ are expressed as $\mathbf{x}_k$, for all $k \in \{1, \ldots, K\}$, such that $\mathbf{x}_k \neq c\mathbf{1}_K$, with c a real constant (and $\mathbf{1}_K$ a column vector of ones of dimension K). The ℓ_1 norm of $\mathbf{x}_k$ is given by $\|\mathbf{x}_k\|_1 = \sum_{i=1}^{N} |x_{ik}|$. The arithmetic means of the variables are given by $\bar{\mathbf{x}}_k$. The Gini Mean Difference between two variables $\mathbf{x}_\ell$ and $\mathbf{x}_k$, proposed by Schechtman and Yitzhaky (1987), is given by :

$$GMD(\mathbf{x}_\ell, \mathbf{x}_k) := \frac{4}{N} \sum_{i=1}^{N} (x_{i\ell} - \bar{\mathbf{x}_\ell})(\hat{F}(x_{ik}) - \bar{F}_{\mathbf{x}_k}), \tag{2.5}$$

where $\hat{F}(x_{ik})$ is the estimated cumulative distribution function of $\mathbf{x}_k$ at point i, $\bar{F}_{\mathbf{x}_k}$ its mean, with $\ell, k = 1, \ldots, K$. When $k = \ell$ the GMD represents the variability of the variable $\mathbf{x}_\ell$ with itself, see Eq.(2.1). Alternatively, it is possible to define the rank vector $R(\mathbf{x}_\ell)$ of variable $\mathbf{x}_\ell$ as an estimator of $F(\mathbf{x}_\ell)$,

$$\hat{F}(x_{i\ell}) = \frac{R(x_{i\ell})}{N} := \begin{cases} \frac{\#\{x \leq x_{i\ell}\}}{N} & \text{if no ties} \\ \frac{\sum_{i=1}^{p} \#\{x \leq x_{i\ell}\}}{Np} & \text{if } p \text{ ties } x_{i\ell}. \end{cases} \tag{2.6}$$

The rank vector assigns the value 1 to the smallest value of vector $\mathbf{x}_\ell$, and so on. In the case of ties, the mean rank is computed as shown below for the first observation :

$$\mathbf{x}_\ell = \begin{pmatrix} 1 \\ 1 \\ 4 \\ 7 \\ 6 \end{pmatrix} \longrightarrow R(\mathbf{x}_\ell) = \begin{pmatrix} 1,5 \\ 1,5 \\ 3 \\ 5 \\ 4 \end{pmatrix} \tag{2.7}$$

A bias corrected estimator of GMD is,

$$GMD(\mathbf{x}_\ell, \mathbf{x}_k) := \frac{4}{N(N-1)} \sum_{i=1}^{N} (x_{i\ell} - \bar{\mathbf{x}}_\ell)(R(x_{ik}) - \bar{R}_{\mathbf{x}_k}), \ \forall k, \ell = 1, \ldots, K, \tag{2.8}$$

with $\bar{R}_{\mathbf{x}_k}$ the mean of the rank vector of variable $\mathbf{x}_k$. The Gini correlation coefficient, the G-correlation frown now on,is defined as follows :

$$GC(\mathbf{x}_\ell, \mathbf{x}_k) := \frac{GMD(\mathbf{x}_\ell, \mathbf{x}_k)}{GMD(\mathbf{x}_\ell, \mathbf{x}_\ell)}, \tag{2.9}$$

with $GC(\mathbf{x}_k, \mathbf{x}_k) = 1$ for all $k = 1, \ldots, K$. Following Yitzhaki (2003), the G-correlation is well-suited for the measurement of correlations in the case of distributions with atypical points and in general in the case of non-normal distributions.

Property 2.2.1. – **Yitzhaki (2003) :**
(i) $-1 \leq GC(\mathbf{x}_\ell, \mathbf{x}_k) \leq 1$.
(ii) *If the variables* $\mathbf{x}_\ell$ *and* $\mathbf{x}_k$ *are independent, for all* $k \neq \ell$*, then* $GC(\mathbf{x}_\ell, \mathbf{x}_k) = GC(\mathbf{x}_k, \mathbf{x}_\ell) = 0$.
(iii) *For any given monotonic transformation* φ, $GC(\mathbf{x}_\ell, \varphi(\mathbf{x}_k)) = GC(\mathbf{x}_\ell, \mathbf{x}_k)$ [*in the same than Spearman's coefficient*].

(iv) *For any given linear transformation* φ, $GC(\varphi(\mathbf{x}_\ell), \mathbf{x}_k) = GC(\mathbf{x}_\ell, \mathbf{x}_k)$ [*in the same manner than Pearson's coefficient*].
(v) *If* $\mathbf{x}_k$ *and* $\mathbf{x}_\ell$ *are exchangeable up to a linear transformation, then* $GC(\mathbf{x}_\ell, \mathbf{x}_k) = GC(\mathbf{x}_k, \mathbf{x}_\ell)$.

Another property could have been added to the previous ones, that of robustness of the G-correlation index. In order to assess its robustness in PCA frameworks, let us define the $K \times K$ G-correlation matrix containing the Gini correlations between each and every pairs of variables :

$$GC(\mathbf{X}) \equiv [GC(\mathbf{x}_\ell, \mathbf{x}_k)]. \tag{2.10}$$

Let $\mathbf{X}^c$ and $\mathbf{R}^c$ be the $N \times K$ column matrices containing respectively, the centered $\mathbf{x}_k$ vectors and the centered rank vectors. On the other hand, let the matrix of basis vectors be $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_h]$ containing the the eigen vector in columns such that there exist h non-zero eigen values $\lambda_1, \ldots, \lambda_h$.

Proposition 2.2.1. *The Gini PCA consists in solving for the eigen values* λ_k *for all* $k = 1, \ldots, K$ *that maximize the Gini variability of* $\boldsymbol{X}$ *such that :*

$$\lambda_k = \arg\max \mathbf{b}_k^\intercal GC(\mathbf{X})\mathbf{b}_k = \arg\max \mathbf{b}_k^\intercal (\mathbf{X}^c)^\intercal \mathbf{R}^c \mathbf{b}_k. \tag{2.11}$$

The eigen values λ_k are derived from the Gini correlation matrix instead of the usual variance-covariance matrix. The eigen vectors are normalized such that $\|\mathbf{b}_k\|_1 = 1$. Then, the observations are projected such that $\mathbf{F} = \mathbf{X}^c\mathbf{B}$, with $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_h]$ the matrix of projected observations into the new subspace spanned by the eigen vectors $\mathbf{b}_k$.

Baccini *and al.* (1996) proposed a ℓ_1 PCA solely based on the diagonal terms of $GC(\mathbf{X})$. In our approach, the extra-diagonal terms, representing the Gini correlation between the variables $\mathbf{x}_k$ and $\mathbf{x}_\ell$ ($k \neq \ell$), are taken into account in order to attenuate the influence of the outliers that could occur in those correlations as well.

2.3 Monte Carlo Simulations

As in standard PCA, the results of the Gini PCA may be interpreted thanks to absolute contributions (ACT) and relative contributions (RCT).

ACT_{ik} represents the share of axis $\mathbf{f}_k$ variability (in the Gini sense) captured by each observation i. This statistics the number of significant components (axis) $\mathbf{f}_k$ to be selected. RCT_{ik} is the distance of observation i towards a component $\mathbf{f}_k$.

Definition 2.3.1. *The absolute contribution of an individual i to the Gini variability of a principal component $\mathbf{f}_k$ is :*

$$ACT_{ik} = \frac{f_{ik} r_{ik}}{GMD(\mathbf{f}_k, \mathbf{f}_k)}, \ \forall k = 1, \ldots, h, \tag{2.12}$$

where r_{ik} is the rank of individual i on the principal axis $\mathbf{f}_k$ and f_{ik} the score of observation i on component $\mathbf{f}_k$.

The absolute contribution of each i to the Gini mean difference of $\mathbf{f}_k$ is such that $ACT_{ik} \in [0, 1]$ and $\sum_{i=1}^{N} ACT_{ik} = 1$.

Definition 2.3.2. *The relative contribution of an individual i to a component $\mathbf{f}_k$ is :*

$$RCT_{ik} = \frac{|f_{ik}|}{\|\mathbf{f}_i\|_1}, \ \forall k = 1, \ldots, h, \tag{2.13}$$

where $\mathbf{f}_i$ is the i-th row of matrix $\mathbf{F}$.

On the one hand, Monte Carlo experiments are conducted with 5-variate normal distributions of size $N = 500$ with independent variables in order to assess the quality of ACT, RCT and λ_k by means of the estimation of Mean Squared Errors (MSE). Let λ_k^{oi} be the eigen value issued from the contamination of the data $\boldsymbol{X}$ by an outlier oi and λ_k the eigen value estimated without contamination. Over 1,000 different possible contaminations, the MSE of λ_k is given by :

$$MSE_{\lambda_k} = \frac{\sum_{i=1}^{1,000} (\lambda_k^{oi} - \lambda_k)^2}{1,000}, \ \forall k = 1, \ldots, h. \tag{2.14}$$

The MSE of ACT et RCT are computed in the same manner.

Algorithm 3: Monte Carlo Simulation

Result: Robust Gini PCA with data contamination

$\theta = 1$ [θ is the value of the outlier] and $N = 500$;

repeat

  Generate a 5-variate normal distribution $\mathbf{X} \sim \mathcal{N}$;

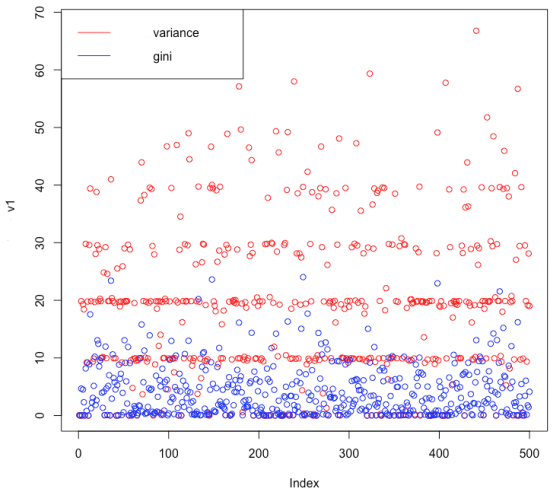
  Contamination : 1 observation (row) of $\mathbf{X}$ is multiplied by θ [random row localization] ;

  Compute ACT^{oi}, RCT^{oi} and λ_k^{oi} ;

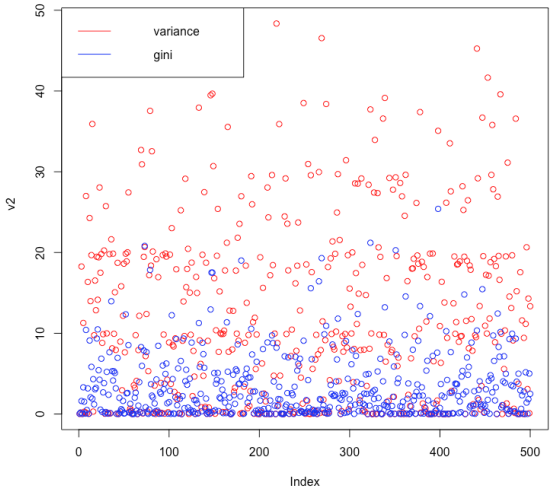
until $\theta = 1,000$ [increment of 1];

return Mean squared Errors of ACT, RCT and λ_k ;

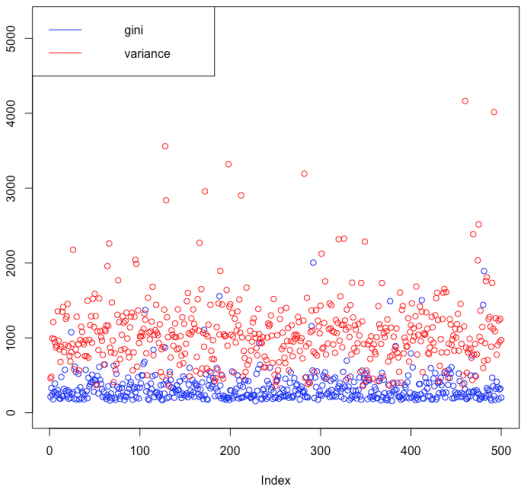
The MSE of the ACT of the 500 observations are computed for components $\mathbf{f}_1$ and $\mathbf{f}_2$, Figure 1a and 1b respectively.



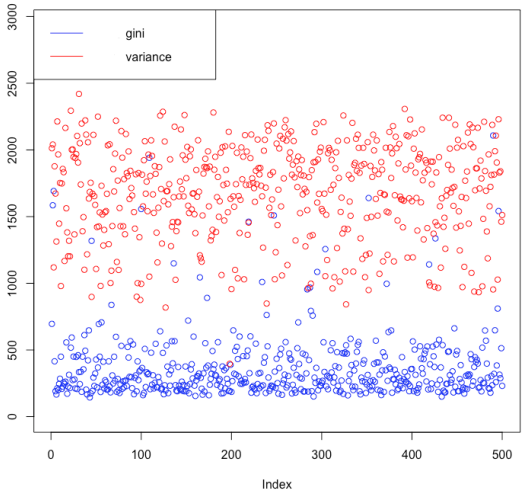
a : ACT_1 Axis 1



b : ACT_2 Axis 2



c : RCT_1 Axis 1



d : RCT_2 Axis 2

FIGURE 2.1 – MSE of the two approaches

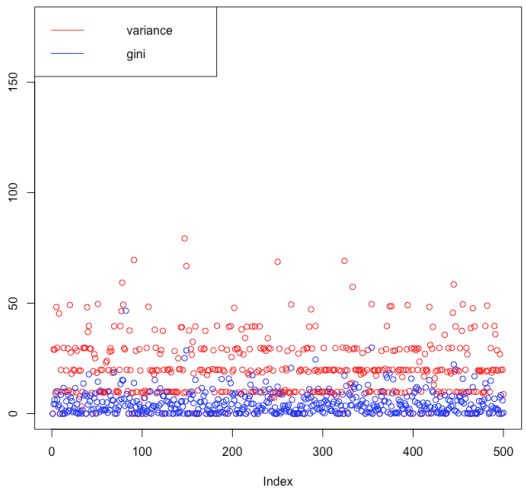
The MSE issued from the Gini PCA (blue points) are less spread out than those of the variance (red points). This means that the quantity of information (dispersion) captured by each observation i remains much more stable with the Gini PCA when the data are increasingly contaminated by θ. The same conclusion holds true for the MSE of RCT (Figures 2.1 c/d).

	% λ_k (Gini)	% λ_k (Var)	MSE (Var)	MSE (Gini)
$axis_1$	0.57	0.90	11.89	0.69
$axis_2$	0.14	0.05	0.79	0.81
$axis_3$	0.11	0.02	0.75	14.15

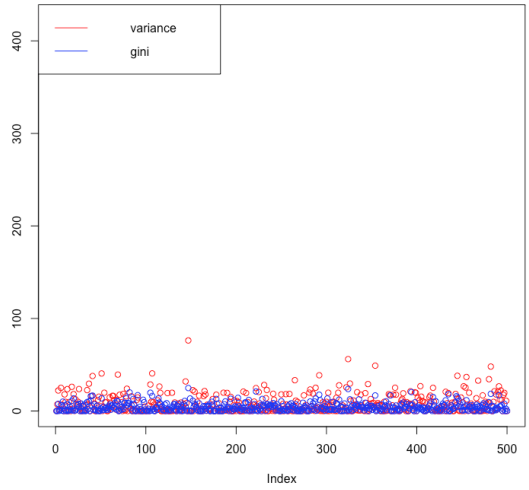
TABLE 2.1 – Eigen Values and MSE : Normal distributions

Table 2.1 below depicts the MSE of the eigen values that are much lower in the Gini PCA (except on axis 3 since the quantity of dispersion is not significant on this axis). The variability on each axis (in mean over 1,000 iterations) $\frac{\lambda_k}{\sum_k \lambda_k} \times 100$ shows that the presence of one outlier drastically affects the repartition of the information on the three components. In the standard PCA (Variance case), the overall variability is important on component 1 (90%), whereas each component must capture $1/5$ of the overall variability (since the 5 variables are independent). The repartition of the variability on each component is more uniform in the Gini case.

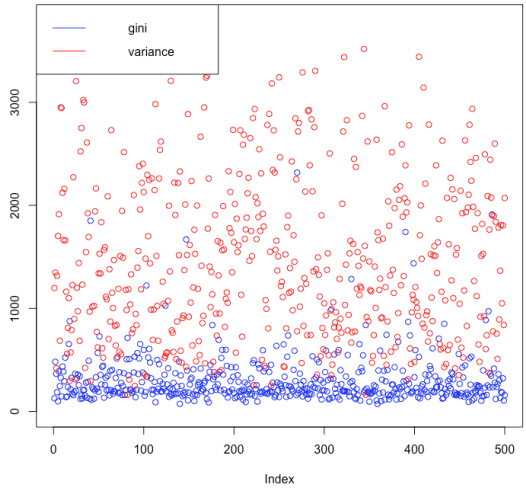
Another Monte Carlo simulation is performed with a mixture of probability distributions of size $N = 500$: Normal [$\mathcal{N}$(0,1)], Gamma [Γ(2,2)], Uniform [$U(\min = 1, \max = 5)$], Cauchy [(location = 0, scale = 1)] and Beta [$\beta(2, 3)$]. The results are similar.



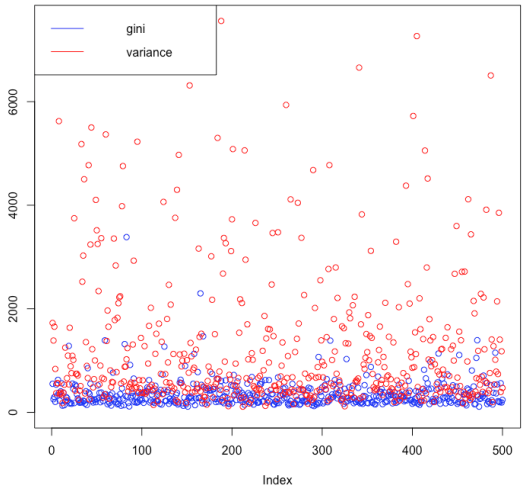
a : ACT_1 Axis 1



b : ACT_2 Axis 2



c : RCT_1 Axis 1



d : RCT_2 Axis 2

FIGURE 2.2 – Absolute and relative Contributions

	% λ_k (Gini)	% λ_k (Var)	MSE (Var)	MSE (Gini)
$axis_1$	0.64	0.88	11.04	0.49
$axis_2$	0.14	0.09	0.87	0.81
$axis_3$	0.09	0.01	0.74	19.25

TABLE 2.2 – Eigen Values and MSE : mixture of distributions

2.4 Concluding remarks

In this chapter, a robust Gini PCA has been performed thanks to the cogini operator underlying the Gini correlation matrix $GC(\boldsymbol{X})$. The interpretations of the Gini PCA, on the basis of ACT, RCT and eigen values, have been shown to be more relevant than the variance case when one outlier affects the sample. This opens the way on using the Gini PCA in many fields. For instance, in financial econometrics, the principal axes are employed as risk factors in order to compute systematic risks and to deduce the risk premium. This also open the way on generalized Gini PCA, which could be based on the generalized cogini operator, see Yitzhaki and Schechtman (2013) and Greselin and Zitikis (2015).

Acknowledgments

This chapter has been presented at the GLADAG 2017 conference in Milan. We would like to thank the organizer, Pr Greselin, and the participants for helpful comments. Remaining errors are ours.

BIBLIOGRAPHIE

A. Baccini, P. Besse and A. de Falguerolles.(1996). *A ℓ_1 norm PCA and a heuristic approach* . Ordinal and Symbolic Data Analysis, E Didday, Y. Lechevalier and O. Opitz (eds), Springer, 359-368.

Banerjee, A. (2010). *A multidimensional Gini index*. Mathematical Social Sciences, **vol. 60,**, pages 87 à 93.

Carcea, M. and Serfling, R. (2015). *A Gini autocovariance function for time series modeling*. Journal of Time Series Analysis, **vol. 7**, 36, 817-838.

Ding, C., Zhou, D., Ha, X., Zha, H. (2006). *R_1-PCA : Rotational Invariant ℓ_1-norm Principal Component Analysis for Robust Subspace Factorization* . Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh.

Gini, C. (1912). *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*, C. Cuppini, Bologna.

Greselin, F. and R. Zitikis. (2015). *Measuring Economic Inequality and Risk : A Unifying Approach Based on Personal Gambles, Societal Preferences and References* . Electronic SSRN Journal.

Olkin, I. and S. Yitzhaki. (1992). *Gini Regression Analysis*. International Statistical Review, **60(2)**, 185-196.

Ouraga, T. (2019). *A note on Gini Principal Component Analysis*.Economics Bulletin, **vol. 39**, Issue 2.

Schechtman, E. & Yitzhaki, S. (2003). *A family of correlation coefficients based on the extended Gini index* . Journal of Economic Inequality, **1(2)**, 129-146.

Yitzhaki, S. & Schechtman, E. (1987). *A Measure of Association Based on Gini's Mean Difference* . Communications in Statistics, **A16**, 207-231.

Yitzhaki, S. (2003). *Gini's Mean difference : a superior measure of variability for non-normal distributions*. METRON - International Journal of Statistics, **vol. LXI**, n. 2, pp. 285 - 316.

Yitzhaki, S. (2003). *TGini's Mean difference : a superior measure of variability for non-normal distributions* . Metron, **LXI(2)**, 285-316.

Yitzhaki, S. & Schechtman, E. (2013). *The Gini Methodology : A Primer on a Statistical Methodology* . New York : Springer.

CHAPITRE 3

PRINCIPAL COMPONENT ANALYSIS : A GENERALIZED GINI APPROACH

Sommaire

Résumé

Une analyse en composantes principales basée sur l'indice de corrélation de Gini généralisé est proposée (ACP-Gini généralisé). L'ACP-Gini généralisée généralise l'ACP-Gini standard. Il est montré, dans le cas gaussien, que l'ACP standard est équivalent à l'ACP-Gini. Il est également prouvé que la réduction de dimensions basée sur la matrice de corrélation de Gini généralisée, qui repose sur les distances Manhattan, est robuste aux valeurs aberrantes. Des simulations de Monte Carlo et une application sur les données des voitures (avec des valeurs aberrantes) montrent la robustesse de l'ACP-Gini généralisé et fournissent différentes interprétations des résultats par rapport à l'ACP basée sur la variance. Ce chapitre est un article en révision, *C.f* Charpentier., Mussard. & Ouraga (2020).

Mots-clés : (R) Statistiques Multivariées ; Gini ; PCA ; Robustesse.

Abstract

A principal component analysis based on the generalized Gini correlation index is proposed (Generalized Gini-PCA). The Generalized Gini-PCA generalizes the standard Gini-PCA. It is shown, in the Gaussian case, that the standard PCA is equivalent to the Gini PCA. It is also proven that the dimensionality reduction based on the generalized Gini correlation matrix, that relies on city-block distances, is robust to outliers. Monte Carlo simulations and an application on cars data (with outliers) show the robustness of the Gini PCA and provide different interpretations of the results compared with the variance PCA. This chapter is a paper in under review, *i.e* Charpentier., Mussard. & Ouraga (2020).

Keywords : (R) Multivariate statistics ; Gini ; PCA ; Robustness.

3.1 Introduction

This late decade, a line of research has been developed and focused on the Gini methodology, see **??** for a general review of different Gini approaches applied in Statistics and in Econometrics. [1] Among the Gini tools, the Gini regression has received a large audience since the Gini regression initiated by Olkin and Yitzhaki (1992). Gini regressions have been generalized by Yitzhaki & Schechtman (2013) in different areas and particularly in time series analysis. Shelef & Schechtman (2011) and Carcea and Serfling (2015) investigated ARMA processes with an identification and an estimation procedure based on Gini autocovariance functions. This robust Gini approach has been shown to be relevant to heavy tailed distributions such as Pareto processes. Also, Shelef (2016) proposed a unit root test based on Gini regressions to deal with outlying observations in the data.

In parallel to the above literature, a second line of research on multidimensional Gini indices arose. This literature paved the way on the valuation of inequality about multiple commodities or dimensions such as education, health, income, etc., that is, to find a real-valued function that quantifies the inequality between the households of a population over each dimension, see among others, List (1999), Gajdos & Weymark (2005), Decancq & Lugo (2013). More recently, Banerjee (2010) shows that it is possible to construct multidimensional Gini indices by exploring the projection of the data in reduced subspaces based on the Euclidean norm. Accordingly, some notions of linear algebra have been increasingly included in the axiomatization of multidimensional Gini indices.

In this chapter, in the same vein as in the second line of research mentioned above, we start from the recognition that linear algebra may be closely related to the maximum level of inequality that arises in a given dimension. In data analysis, the variance maximization is mainly used to further analyze projected data in reduced subspaces. The variance criterion implies many problems since it captures a very precise notion of dispersion, which does not always match some basic properties satisfied by variability measures such as the Gini index. Such a property may be, for example, an invariance condition postulating that a dispersion measure remains

1. See Giorgi (2013) for an overview of the "Gini methodology".

constant when the data are transformed by monotonic maps. [2] Another property typically related to the Gini index is its robustness to outlying observations, see e.g. Yitzhaki & Olkin (1991) in the case of linear regressions. Accordingly, it seems natural to analyze multidimensional dispersion with the Gini index, instead of the variance, in order to provide a Principal Components Analysis (PCA) in a Gini sense (Gini PCA). [3]

In the field of PCA, the ℓ_1 norm PCA became famous these last two decades. The algorithms underlying the ℓ_1 norm PCA are mainly built on the minimization of the absolute difference between the coordinates of the projected data and the original data.

Kwak (2008), show that the ℓ_1 norm PCA is rather less sensitive to outliers compared with the ℓ_2-norm PCA. An algorithm based on either linear or quadratic programming is proposed to obtain a ℓ_1 norm PCA. Despite the robustness of the technique, it remains quite time-consuming and not rotational invariant.

Ding *et al.* (2006) propose the R1-PCA which is a rotational invariant $\ell_1$norm PCA. It deals properly with outliers and enables rotations, however it strongly depends on the dimension of the subspace in which the data are projected onto. Indeed, the projector of dimension $K-1$ cannot be deduced from the projector of dimension K.

Since the ℓ_1 norm PCA relies on optimization problems, the literature offers some new techniques of optimization related to the ℓ_1 PCA such as ℓ_1 norm discriminant analysis for image and pattern recognitions, see for instance Li *et. al* (2015) who propose to replace the Euclidean norm by the ell_1 norm in order to maximize the between-group variability in a given sample with the aid of an iterative algorithm (see also, Brooks *et al.* (2013) for the ℓ_1 norm best-fit hyperplane problem leading to a 'pure' ℓ_1 PCA).

Instead of looking for optimization procedures related to the ℓ_1 norm, we propose a closed-form ℓ_1 norm PCA. [4] Indeed, following the recent

2. See Furman & Zitikis (2017) for the link between variability (risk) measures and the Gini correlation index. See also Laurini & Ohashi (2015) for problem of noisy PCA for pricing interest rate derivatives.

3. Recent PCAs derive latent variables thanks to regressions based on elastic net (a ℓ_1 regularization) that improves the quality of the regression curve estimation, see Zou, Hastie & Tibshirani (2006).

4. For ℓ_1 norm PCA with new geometric perspectives, see Schölkopf *et al.* (1998) for the kernel PCA which is a non-linear PCA.

works on the Gini index done by Yitzhaki & Schechtman (2013), it is well-known that the Gini covariance function, being a ℓ_1 norm covariance, enables robustness statistical estimations. Baccini, Besse & de Falguerolles (1996) and Korhonen & Siljamäki (1998) are among the first authors dealing with a ℓ_1-norm PCA Gini framework. Their idea was to robustify the standard PCA by means of the Gini Mean Difference metric introduced by Gini (1912), which is a city-block distance measure of variability. The authors, *i.e* Charpentier, Mussard and Ouraga (2020) employ the Gini Mean Difference as an estimator of the standard deviation of each variable before running the singular value decomposition leading to a robust PCA. In what follows, we investigate the employ of the generalized Gini covariance opertator in order to obtained a closed-form ℓ_1 PCA without taking recourse to algorithms of optimization.

In particular, it is shown that the variance may be seen as an inappropriate criterion for dimensionality reduction in the case of data contamination or outlying observations. A generalized Gini PCA is investigated by means of Gini correlations matrices. These matrices contain generalized Gini correlation coefficients (see Yitzhaki (2003)) based on the Gini covariance operator introduced by Schechtman & Yitzhaki (1987) and Yitzhaki & Schechtman (2003). The generalized Gini correlation coefficients are : (i) bounded, (ii) invariant to monotonic transformations, (iii) and symmetric whenever the variables are exchangeable. It is shown that the standard PCA is equivalent to the Gini PCA when the variables are Gaussians. Also, it is shown that the generalized Gini PCA may be realized either in the space of the variables or in the space of the observations. In each case, some statistics are proposed to perform some interpretations of the variables and of the observations (absolute and relative contributions). To be precise, an U-statistics test is introduced to test for the significance of the correlations between the axes of the new subspace and the variables in order to assess their significance. Monte Carlo simulations are performed in order to show the superiority of the Gini PCA compared with the usual PCA when outlying observations contaminate the data. Finally, with the aid of the well-known cars data, which contain outliers, it is shown that the generalized Gini PCA leads to different results compared with the usual PCA.

The outline of the paper is as follows. Section 3.2 sets the notations and

presents some ℓ_2 norm approaches of PCA. Section 3.3 reviews the Gini-covariance operator. Section 3.4 is devoted to the generalized Gini PCA. Section 3.5 focuses on the interpretation of the Gini PCA. Sections 3.6 and 3.7 present some Monte Carlo simulations and applications, respectively.

3.2 Motivations for the use of Gini PCA

In this Section, the notations and the assumptions are set. Since our ℓ_1-norm PCA relies on closed form PCA inspired from ℓ_2-norm PCAs, we briefly review some ℓ_2-norm PCAs in order to motivate the employ of the Gini PCA.

Notations and definitions

Let $\mathbb{N}^*$ be the set of integers and $\mathbb{R}$ $[\mathbb{R}_{++}]$ the set of [positive] real numbers. Let $\mathcal{M}$ be the set of all $N \times K$ matrix $\boldsymbol{X} = [x_{ik}]$ that describes N observations on K dimensions such that $N \gg K$, with elements $x_{ik} \in \mathbb{R}$, and $\mathbb{I}_n$ the $n \times n$ identity matrix. The $N \times 1$ vectors representing each variable are expressed as $\mathbf{x}_{\cdot k}$, for all $k \in \{1, \ldots, K\}$ and we assume that $\mathbf{x}_{\cdot k} \neq c\mathbf{1}_N$, with c a real constant and $\mathbf{1}_N$ a N-dimensional column vector of ones. The $K \times 1$ vectors representing each observation i (the transposed ith line of $\boldsymbol{X}$) are expressed as $\mathbf{x}_{i\cdot}$, for all $i \in \{1, \ldots, N\}$. It is assumed that $\mathbf{x}_{\cdot k}$ is the realization of the random variable X_k, with cumulative distribution function F_k. The arithmetic mean of each column (line) of the matrix $\boldsymbol{X}$ is given by $\bar{\mathbf{x}}_{\cdot k}$ ($\bar{\mathbf{x}}_{i\cdot}$). The cardinal of set A is denoted $\#\{A\}$. The ℓ_1 norm, for any given real vector $\mathbf{x}$, is $\|\mathbf{x}\|_1 = \sum_{k=1}^{K} |x_{\cdot k}|$, whereas the ℓ_2 norm is $\|\mathbf{x}\|_2 = (\sum_{k=1}^{K} x_{\cdot k}^2)^{1/2}$.

Assumption 3.2.1. *The random variables X_k are such that $\mathbb{E}[|X_k|] < \infty$ for all $k \in \{1, \ldots, K\}$, but no assumption is made on the second moments (that may not exist).*

This assumption imposes less structure compared with the classical PCA in which the existence of the second moments are necessary, as can be seen in the next subsection.

Variants of PCA based on the ℓ_2 norm

The classical formulation of the PCA, to obtain the first component, can be obtained by solving

$$\omega_1^* \in \text{argmax} \left\{ \text{Var}[\boldsymbol{X}\omega] \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1, \tag{3.1}$$

or equivalently

$$\omega_1^* \in \text{argmax} \left\{ \omega^\top \boldsymbol{\Sigma} \omega \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1, \tag{3.2}$$

where $\omega \in \mathbb{R}^K$, and $\boldsymbol{\Sigma}$ is the (symmetric positive semi-definite) $K \times K$ sample covariance matrix. Mardia, Kent & Bibby (1979) suggest to write

$$\omega_1^* \in \text{argmax} \left\{ \sum_{j=1}^{K} \text{Var}[\mathbf{x}_{\cdot,j}] \cdot \text{Cor}[\mathbf{x}_{\cdot,j}, \boldsymbol{X}\omega] \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1.$$

With scaled variables [5] (i.e. $\text{Var}[\mathbf{x}_{\cdot,j}] = 1, \forall j$)

$$\omega_1^* \in \text{argmax} \left\{ \sum_{j=1}^{K} \text{Cor}[\mathbf{x}_{\cdot,j}, \boldsymbol{X}\omega] \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1. \tag{3.3}$$

Then a Principal Component Pursuit can start : we consider the 'residuals', $\boldsymbol{X}_{(1)} = \boldsymbol{X} - \boldsymbol{X}\omega_1^* \omega_1^{*\top}$, its covariance matrix $\boldsymbol{\Sigma}_{(1)}$, and we solve

$$\omega_2^* \in \text{argmax} \left\{ \omega^\top \boldsymbol{\Sigma}_{(1)} \omega \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1.$$

The part $\boldsymbol{X}\omega_1^* \omega_1^{*\top}$ is actually a constraint that we add to ensure the orthogonality of the two first components. This problem is equivalent to finding the maxima of $\text{Var}[\boldsymbol{X}\omega]$ subject to $\|\omega\|_2^2 = 1$ and $\omega \perp \omega_1^*$. This idea is also called Hotelling (or Wielandt) deflation technique. On the k-th iteration, we extract the leading eigenvector

$$\omega_k^* \in \text{argmax} \left\{ \omega^\top \boldsymbol{\Sigma}_{(k-1)} \omega \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1,$$

5. In most cases, PCA is performed on scaled (and centered) variables, otherwise variables with large scales might alter interpretations. Thus, it will make sense, later on, to assume that components of $\boldsymbol{X}$ have identical distributions. At least the first two moments will be equal.

where $\mathbf{\Sigma}_{(k-1)} = \mathbf{\Sigma}_{(k-2)} - \omega^*_{k-1}\omega^{*\top}_{k-1}\mathbf{\Sigma}_{(k-1)}\omega^*_{k-1}\omega^{*\top}_{k-1}$ (see e.g. Saad (1998)). Note that, following Hotelling (1933) and Eckart & Young (1936), that it is also possible to write this problem as

$$\min\left\{\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_\star\right\} \text{ subject to } \text{rank}[\tilde{\boldsymbol{X}}] \leq k$$

where $\|\cdot\|_\star$ denotes the nuclear norm of a matrix (*i.e.* the sum of its singular values) [6].

One extension, introduced in d'Aspremont *et al.* (1920), was to add a constraint based on the cardinality of ω (also called ℓ_0 norm) corresponding to the number of non-zero coefficients of ω. The penalized objective function is then

$$\max\left\{\omega^\top\mathbf{\Sigma}\omega - \lambda\text{card}[\omega]\right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top\omega = 1,$$

for some $\lambda > 0$. This is called *sparse PCA*, and can be related to sparse regression, introduced in Tibshirani (1996). But as pointed out in Mackey (2009), interpretation is not easy and the components obtained are not orthogonal. Gorban *et al.* (2007) considered an extension to nonlinear Principal Manifolds to take into account nonlinearities.

Another direction for extensions was to consider Robust Principal Component Analysis. Candes *et al.* (2009) suggested an approach based on the fact that principal component pursuit can be obtained by solving

$$\min\left\{\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_\star + \lambda\|\tilde{\boldsymbol{X}}\|_1\right\}.$$

But other methods were also considered to obtain Robust PCA. A natural 'scale-free' version is obtained by considering a rank matrix instead of $\boldsymbol{X}$. This is also called 'ordinal' PCA in the literature, see Korhonen & Siljamäki (1998). The first 'ordinal' component is

$$\omega_1^* \in \text{argmax}\left\{\sum_{j=1}^{K}\mathcal{R}[\mathbf{x}_{\cdot,j}, \boldsymbol{X}\omega]\right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top\omega = 1 \tag{3.4}$$

where $\mathcal{R}$ denotes some rank based correlation, *e.g.* Spearman's rank correlation, as an extention of Equation (3.3). So, quite naturally, one possible

6. but other norms have also been considered in statistical literature, such as the Froebenius norm in the Eckart-Young theorem, or the maximum of singular values – also called 2-(induced)-norm.

extension of Equation (3.2) would be

$$\omega_1^* \in \text{argmax}\left\{\omega^\top \mathcal{R}[\boldsymbol{X}]\omega\right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top\omega = 1$$

where $\mathcal{R}[\boldsymbol{X}]$ denotes Spearman's rank correlation. In this section, instead of using Pearson's correlation (as in Equation (3.2) when the variables are scaled) or Spearman's (as in this ordinal PCA), we will consider the multidimensional Gini correlation based on the h-covariance operator.

3.3 Geometry of Gini PCA : Gini-Covariance Operators

The first PCA was introduced by Pearson (1901), projecting $\boldsymbol{X}$ onto the eigenvectors of its covariance matrix, and observing that the variances of those projections are the corresponding eigenvalues. One of the key property is that $\boldsymbol{X}^\top\boldsymbol{X}$ is a positive matrix. Most statistical properties of PCAs (see Flury & Riedwyl (1988) or Anderson (1963)) are obtained under Gaussian assumptions. Furthermore, geometric properties can be obtained using the fact that the covariance defines an inner product on the subspace of random variables with finite second moment (up to a translation, *i.e.* we identify any two that differ by a constant).

We will discuss in this section the properties of the Gini Covariance operator with the special case of Gaussian random variables, and the property of the Gini correlation matrix that will be used in the next Section for the Gini PCA.

The Gini-covariance operator

In this section, $\boldsymbol{X} = (X_1, \cdots, X_K)$ denotes a random vector. The covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$, two random vectors, is defined as the inner product between centered versions of the vectors,

$$\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^\intercal]. \tag{3.5}$$

Hence, it is the matrix where elements are regular covariances between components of the vectors, $\text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = [\text{Cov}(X_i, Y_j)]$. It is the upper-right block of the covariance matrix of $(\boldsymbol{X}, \boldsymbol{Y})$. Note that $\text{Cov}(\boldsymbol{X}, \boldsymbol{X})$ is the standard variance-covariance matrix of vector $\boldsymbol{X}$.

Definition 3.3.1. *Let $\boldsymbol{X} = (X_1, \cdots, X_K)$ be collections of K identically distributed random variables. Let $h : \mathbb{R} \to \mathbb{R}$ denote a non-decreasing function. Let $h(\boldsymbol{X})$ denote the random vector $(h(X_1), \cdots, h(X_K))$, and assume that each component has a finite variance. Then, operator $\Gamma C_h(\boldsymbol{X}) = \text{Cov}(\boldsymbol{X}, h(\boldsymbol{X}))$ is called h-Gini covariance matrix.*

Since h is a non-decreasing mapping, then $\boldsymbol{X}$ and $h(\boldsymbol{X})$ are componentwise comonotonic random vectors. Assuming that components of $\boldsymbol{X}$ are identically distributed is a reasonable assumption in the context of scaled (and centered) PCA, as discussed in footnote 5. Nevertheless, a stronger technical assumption will be necessary : pairwise-exchangeability.

Definition 3.3.2. *$\boldsymbol{X}$ is said to be pairwise-exchangeable if for all pair $(i, j) \in \{1, \cdots, K\}^2$, (X_i, X_j) is exchangeable, in the sense that $(X_i, X_j) \overset{\mathcal{L}}{=} (X_j, X_i)$.*

Pairwise-exchangeability is a stronger concept than having only one vector with identically distributed components, and a weaker concept than (full) exchangeability. In the Gaussian case where $h(X_k) = \Phi(X_k)$ with $\Phi(X_k)$ being the normal cdf of X_k for all $k = 1, \ldots, K$, pairwise-exchangeability is equivalent to components identically distributed.

Proposition 3.3.1. *If $\boldsymbol{X}$ is a Gaussian vector with identically distributed components, then $\boldsymbol{X}$ is pairwise-exchangeable.*

Démonstration. For simplicity, assume that components of $\boldsymbol{X}$ are $\mathcal{N}(0, 1)$ random variables, then $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\rho})$ where $\boldsymbol{\rho}$ is a correlation matrix. In that case

$$\begin{bmatrix} X_i \\ X_j \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ij} \\ \rho_{ji} & 1 \end{bmatrix} \right),$$

with Pearson correlation $\rho_{ij} = \rho_{ji}$, thus (X_i, X_j) is exchangeable. □

Let us now introduce the Gini-covariance. Gini (1912) introduced the Gini mean difference operator Δ, defined as :

$$\Delta(X) = \mathbb{E}(|X_1 - X_2|) \text{ where } X_1, X_2 \sim X, \text{ and } X_1 \perp\!\!\!\perp X_2, \tag{3.6}$$

for some random variable X (or more specifically for some distribution F with $X \sim F$, because this operator is law invariant). One can rewrite :

$$\Delta(X) = 4\text{Cov}[X, F(X)] = \frac{1}{3} \frac{\text{Cov}[X, F(X)]}{\text{Cov}[F(X), F(X)]}$$

where the term on the right is interpreted as the slope of the regression curve of the observed variable X and its 'ranks' (up to a scaling coefficient). Thus, the Gini-covariance is obtained when the function h is equal to the cumulative distribution function of the second term, see Schechtman & Yitzhaki (1987).

Definition 3.3.3. *Let $\boldsymbol{X} = (X_1, \cdots, X_K)$ be a collection of K identically distributed random variables, with cumulative distribution function F. Then, the Gini covariance is $\Gamma C_F(\boldsymbol{X}) = \text{Cov}(\boldsymbol{X}, F(\boldsymbol{X}))$.*

On this basis, it is possible to show that the Gini covariance matrix is a positive semi-definite matrix.

Theorem 3.3.1. *Let $Z \sim \mathcal{N}(0,1)$. If $\boldsymbol{X}$ represents identically distributed Gaussian random variables, with distribution $\mathcal{N}(\mu, \sigma^2)$, then the two following assertions hold :*
(i) $\Gamma C_F(\boldsymbol{X}) = \sigma^{-1}\text{Cov}(Z, \Phi(Z))\text{Var}(\boldsymbol{X})$.
(ii) $\Gamma C_F(\boldsymbol{X})$ *is a positive-semi definite matrix.*

Démonstration. (i) In the Gaussian case, if h is the cumulative distribution function of the X_k's, then $\text{Cov}(X_k, h(X_\ell)) = r\sigma \cdot \text{Cov}(Z, \Phi(Z))$, where Φ is the normal cdf, see Yitzhaki & Schechtman (2013), Chapter 3. Observe that $\text{Cov}(X_k, h(X_k)) = \sigma \cdot \text{Cov}(Z, \Phi(Z))$, if h is the cdf of X_k. Thus, $\lambda := cov(Z, \Phi(Z))$ yields :

$$\Gamma C_F((X_k, X_\ell)) = \lambda \begin{bmatrix} \sigma & \rho\sigma \\ \rho\sigma & \sigma \end{bmatrix} = \lambda\sigma \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \frac{\lambda}{\sigma}\text{Var}((X_k, X_\ell)).$$

(ii) We have $\text{Cov}(Z, \Phi(Z)) \geq 0$, then it follows that $C_F((X_k, X_\ell)) \geq 0$:

$$\mathbf{x}^\top C_F(\boldsymbol{X})\mathbf{x} = \mathbf{x}^\top \frac{cov(Z, \Phi(Z))}{\sigma}\text{Var}(\boldsymbol{X})\mathbf{x} \geq 0,$$

which ends the proof. □

Note that $\Gamma C_F(\boldsymbol{X}) = \text{Cov}(\boldsymbol{X}, -\overline{F}(\boldsymbol{X})) = \Gamma C_{-\overline{F}}(\boldsymbol{X})$, where $\overline{F}$ denotes the survival distribution function.

Definition 3.3.4. *Let $\boldsymbol{X} = (X_1, \cdots, X_K)$ be a collection of K identically distributed random variables, with survival distribution function $\overline{F}$. Then, the generalized Gini covariance is $G\Gamma C_\nu(\boldsymbol{X}) = \Gamma C_{-\overline{F}^{\nu-1}}(\boldsymbol{X}) = \text{Cov}(\boldsymbol{X}, -\overline{F}^{\nu-1}(\boldsymbol{X}))$, for $\nu > 1$.*

This operator is related to the one introduced in Yitzhaki & Schechtman (2003), called generalized Gini mean difference GMD_ν operator. More precisely, an estimator of the generalized Gini mean difference is given by :

$$GMD_\nu(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) := -\frac{2}{N-1}\nu cov(\mathbf{x}_{.\ell}, \mathbf{R}^{\nu-1}_{\mathbf{x}_{.k}}),\ \nu > 1,$$

where $\mathbf{R}_{\mathbf{x}_{.k}} = (R(x_{1k}), \ldots, R(x_{nk}))$ is the decumulative rank vector of $\mathbf{x}_{.k}$, that is, the vector that assigns the smallest value (1) to the greatest observation x_{ik}, and so on. The rank of observation i with respect to variable k is :

$$R(x_{ik}) := \begin{cases} N+1-\#\{x \leq x_{ik}\} & \text{if no ties} \\ N+1-\frac{1}{p}\sum_{i=1}^{p}\#\{x \leq x_{ik}\} & \text{if } p \text{ ties } x_{ik}. \end{cases}$$

Hence $GMD_\nu(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})$ is the empirical version of

$$2\nu\Gamma C_\nu(X_\ell, X_k) := -2\nu cov\Big(X_\ell, \overline{F}_k(X_k)^{\nu-1}\Big).$$

The index GMD_ν is a generalized version of the GMD_2 proposed earlier by Schechtman & Yitzhaki (1987), and can also be written as :

$$GMD_2(X_k, X_k) = 4cov\Big(X_k, F_k(X_k)\Big) = \Delta(X_k).$$

When $k = \ell$, GMD_ν represents the variability of the variable $\mathbf{x}_{.k}$ itself. Focus is put on the lower tail of the distribution $\mathbf{x}_{.k}$ whenever $\nu \to \infty$, the approach is said to be max-min in the sense that GMD_ν inflates the minimum value of the distribution. On the contrary, whenever $\nu \to 0$, the approach is said to be max-max, in this case focus is put on the upper tail of the distribution $\mathbf{x}_{.k}$. As mentioned in Yitzhaki & Schechtman (2013), the case $\nu < 1$ does not entail simple interpretations, thereby the parameter ν is used to be set as $\nu > 1$ in empirical applications. [7]

Note that even if X_k and X_ℓ have the same distribution, we might have $GMD_\nu(X_k, X_\ell) \neq GMD_\nu(X_\ell, X_k)$, as shown on the example of Figure 3.1. In that case $\mathbb{E}[X_k h(X_\ell)] \neq \mathbb{E}[X_\ell h(X_k)]$ if $h(2) \neq 2h(1)$ (this property is nevertheless valid if h is linear). We would have $GMD_\nu(X_k, X_\ell) = GMD_\nu(X_\ell, X_k)$ when X_k and X_ℓ are exchangeable. But since generally GMD_ν is not symmetric, we have for $\mathbf{x}_{.k}$ being not a monotonic transformation of $\mathbf{x}_{.\ell}$ and $\nu > 1$, $GMD_\nu(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}) \neq GMD_\nu(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})$.

7. In risk analysis $\nu \in (0, 1)$ denotes risk lover decision makers (max-max approach), whereas $\nu > 1$ stands for risk averse decision makers, and $\nu \to \infty$ extreme risk aversion (max-min approach).

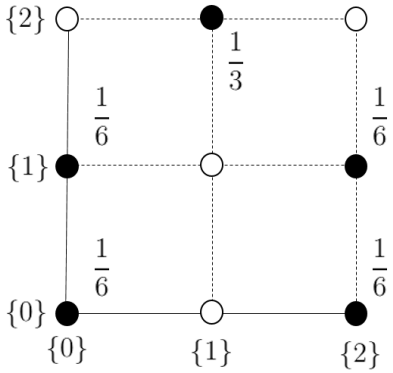


FIGURE 3.1 – Joint distribution of a random pair (X_k, X_ℓ) such that $\mathbb{E}[X_k h(X_\ell)] \neq \mathbb{E}[X_\ell h(X_k)]$, with non-exchangeable components $X_k \overset{\mathcal{L}}{=} X_\ell$.

Generalized Gini correlation

In this section, $\boldsymbol{X}$ is a matrix in $\mathcal{M}$. The Gini correlation coefficient (G-correlation frown now on), is a normalized GMD_ν index such that for all $\nu > 1$, see Yitzhaki & Schechtman (2003),

$$GC_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k}) := \frac{GMD_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k})}{GMD_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot\ell})} \; ; \; GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot\ell}) := \frac{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot\ell})}{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k})},$$

with $GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k}) = 1$ and $GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k}) \neq 0$, for all $k, \ell = 1, \ldots, K$. Following Yitzhaki & Schechtman (2003), the G-correlation is well-suited for the measurement of correlations between non-normal distributions or in the presence of outlying observations in the sample.

Property 3.3.1. – **Schechtman and Yitzhaki (2013) :**
(i) $GC_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k}) \leq 1$.
(ii) *If the variables* $\mathbf{x}_{\cdot\ell}$ *and* $\mathbf{x}_{\cdot k}$ *are independent, for all* $k \neq \ell$, *then* $GC_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k}) = GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot\ell}) = 0$.
(iii) *For any given monotonic increasing transformation* φ, $GC_\nu(\mathbf{x}_{\cdot\ell}, \varphi(\mathbf{x}_{\cdot k})) = GC_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k})$.
(iv) *If* $(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k})$ *have a bivariate normal distribution with Pearson correlation* ρ, *then* $GC_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k}) = GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot\ell}) = \rho$.
(v) *If* $\mathbf{x}_{\cdot k}$ *and* $\mathbf{x}_{\cdot\ell}$ *are exchangeable up to a linear transformation, then* $GC_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k}) = GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot\ell})$.

Whenever $\nu \to 1$, the variability of the variables is attenuated so that GMD_ν tends to zero (even if the variables exhibit a strong variance). The choice of ν is interesting to perform generalized Gini PCA with various values of ν in order to robustify the results of the PCA, since the standard PCA (based on the variance) is potentially of bad quality if outlying observations drastically affect the sample.

A G-correlation matrix is proposed to analyze the data into a new vector space. Following Property 3.3.1 (iv), it is possible to rescale the variables $\mathbf{x}_{.\ell}$ thanks to a linear transformation, then the matrix of standardized observation is,

$$\boldsymbol{Z} \equiv [z_{i\ell}] := \left[\frac{x_{i\ell} - \bar{\mathbf{x}}_{.\ell}}{GMD_\nu(\mathbf{x}_{.\ell}, \mathbf{x}_{.\ell})} \right]. \tag{3.7}$$

The variable $z_{i\ell}$ is a real number without dimension. The variables $\mathbf{x}_{.k}$ are rescaled such that their Gini variability is equal to unity. Now, we define the $N \times K$ matrix of decumulative centered rank vectors of $\boldsymbol{Z}$, which are the same compared with those of $\boldsymbol{X}$:

$$\boldsymbol{R}^c_{\mathbf{z}} \equiv [R^c(z_{i\ell})] := [R(z_{i\ell})^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}_{\mathbf{z}_{.\ell}}] = [R(x_{i\ell})^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}_{\mathbf{x}_{.\ell}}].$$

Note that the last equality holds since the standardization (5.2) is a strictly increasing affine transformation. [8] The $K \times K$ matrix containing all G-correlation indices between all couples of variables $\mathbf{z}_{.k}$ and $\mathbf{z}_{.\ell}$, for all $k, \ell = 1, \ldots, K$ is expressed as :

$$GC_\nu(\boldsymbol{Z}) := -\frac{2\nu}{N(N-1)} \boldsymbol{Z}^\top \boldsymbol{R}^c_{\mathbf{z}}.$$

Indeed, if $GMD_\nu(\boldsymbol{Z}) \equiv [GMD_\nu(\mathbf{z}_{.k}, \mathbf{z}_{.\ell})]$, then we get the following.

Proposition 3.3.2. *For each standardized matrix $\boldsymbol{Z}$ defined in (5.2), the following relations hold :*

$$GMD_\nu(\boldsymbol{Z}) = GC_\nu(\boldsymbol{X}) = GC_\nu(\boldsymbol{Z}). \tag{3.8}$$

$$GMD_\nu(\mathbf{z}_{.k}, \mathbf{z}_{.k}) = 1, \ \forall k = 1, \ldots, K. \tag{3.9}$$

8. By definition $GMD_\nu(\mathbf{x}_{.\ell}, \mathbf{x}_{.\ell}) \geq 0$ for all $\ell = 1, \ldots, K$. As we impose that $\mathbf{x}_{.\ell} \neq c\mathbf{1}_N$, the condition becomes $GMD_\nu(\mathbf{x}_{.\ell}, \mathbf{x}_{.\ell}) > 0$.

Démonstration. We have $GMD_\nu(\boldsymbol{Z}) \equiv [GMD_\nu(\mathbf{z}_{\cdot k}, \mathbf{z}_{\cdot \ell})]$ being a $K \times K$ matrix. The extra diagonal terms may be rewritten as,

$$\begin{aligned} GMD_\nu(\mathbf{z}_{\cdot k}, \mathbf{z}_{\cdot \ell}) &= -\frac{2}{N-1} \nu cov(\mathbf{z}_{\cdot k}, \mathbf{R}_{\mathbf{z}_{\cdot \ell}}^{\nu-1}) \\ &= -\frac{2}{N-1} \nu cov\left(\frac{\mathbf{x}_{\cdot k} - \bar{\mathbf{x}}_{\cdot k} \mathbf{1}_N}{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k})}, \mathbf{R}_{\mathbf{z}_{\cdot \ell}}^{\nu-1} \right) \\ &= -\frac{2}{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k})} \left[\frac{\nu cov(\mathbf{x}_{\cdot k}, \mathbf{R}_{\mathbf{z}_{\cdot \ell}}^{\nu-1})}{N-1} - \frac{\nu cov(\bar{\mathbf{x}}_{\cdot k} \mathbf{1}_N, \mathbf{R}_{\mathbf{z}_{\cdot \ell}}^{\nu-1})}{N-1} \right] \\ &= \frac{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot \ell})}{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k})} = GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot \ell}). \end{aligned}$$

Finally, using the same approach as before, we get :

$$\begin{aligned} GMD_\nu(\mathbf{z}_{\cdot k}, \mathbf{z}_{\cdot k}) &= -\frac{2}{N-1} \nu cov(\mathbf{z}_{\cdot k}, \mathbf{R}_{\mathbf{z}_{\cdot k}}^{\nu-1}) \\ &= \frac{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k})}{GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k})} \\ &= GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot k}) = 1. \end{aligned}$$

By Property 3.3.2 (iv), since $\mathbf{R}_{\mathbf{x}_{\cdot k}} = \mathbf{R}_{\mathbf{z}_{\cdot k}}$, then $GC_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot \ell}) = GC_\nu(\mathbf{z}_{\cdot k}, \mathbf{z}_{\cdot \ell})$. Thus,

$$GMD_\nu(\boldsymbol{Z}) = -\frac{2\nu}{N(N-1)} \boldsymbol{Z}^\top \boldsymbol{R}_{\mathbf{z}}^c = GC_\nu(\boldsymbol{X}) = GC_\nu(\boldsymbol{Z}),$$

which ends the proof. □

Finally, under a normality assumption, the generalized Gini covariance matrix $GC_\nu(\boldsymbol{X}) \equiv [GMD_\nu(X_k, X_\ell)]$ is shown to be a positive semi-definite matrix.

Theorem 3.3.2. *Let $Z \sim \mathcal{N}(0,1)$. If $\boldsymbol{X}$ represents identically distributed Gaussian random variables, with distribution $\mathcal{N}(\mu, \sigma^2)$, then the two following assertions holds :*

(i) $GC_\nu(\boldsymbol{X}) = \sigma^{-1}\text{Cov}(Z, \Phi(Z))\text{Var}(\boldsymbol{X})$.

(ii) $GC_\nu(\boldsymbol{X})$ *is a positive semi-definite matrix.*

Démonstration. The first part (i) follows from Yitzhaki & Schechtman (2013), Chapter 6. The second part follows directly from (i). □

Theorem 3.3.2 shows that under the normality assumption, the variance is a special case of the Gini methodology. As a consequence, for multivariate normal distributions, it is shown in Section 3.4 that Gini PCAs and classical PCA (based on the ℓ_2 norm and the covariance matrix) are equivalent.

3.4 Generalized Gini PCA

In this section, the multidimensional Gini variability of the observations $i = 1, \ldots, N$, embodied by the matrix $GC_\nu(\boldsymbol{Z})$, is maximized in the $\mathbb{R}^K$-Euclidean space, *i.e.*, in the set of variables $\{\mathbf{z}_{\cdot 1}, \ldots, \mathbf{z}_{\cdot K}\}$. This allows the observations to be projected onto the new vector space spanned by the eigenvectors of $GC_\nu(\boldsymbol{Z})$. Then, the projection of the variables is investigated in the $\mathbb{R}^N$-Euclidean space induced by $GC_\nu(\boldsymbol{Z})$. Both observations and variables are analyzed through the prism of *absolute* and *relative* contributions to propose relevant interpretations of the data in each subspace.

The $\mathbb{R}^K$-Euclidean space

It is possible to investigate the projection of the data $\boldsymbol{Z}$ onto the new vector space induced by $GMD_\nu(\boldsymbol{Z})$ or alternatively by $GC_\nu(\boldsymbol{Z})$ since $GMD_\nu(\boldsymbol{Z}) = GC_\nu(\boldsymbol{Z})$. Let $\mathbf{f}_{\cdot k}$ be the kth principal component, *i.e.* the kth axis of the new subspace, such that the $N \times K$ matrix $\boldsymbol{F}$ is defined by $\boldsymbol{F} \equiv [\mathbf{f}_{\cdot 1}, \ldots, \mathbf{f}_{\cdot K}]$ with $\boldsymbol{R}^c_{\mathbf{f}} \equiv [\mathbf{R}^{\nu-1}_{c,\mathbf{f}_1}, \ldots, \mathbf{R}^c_{c,\mathbf{f}_{\cdot K}}]$ its corresponding decumulative centered rank matrix (where each decumulative rank vector is raised to an exponent of $\nu - 1$). The $K \times K$ matrix $\boldsymbol{B} \equiv [\mathbf{b}_{\cdot 1}, \ldots, \mathbf{b}_{\cdot K}]$ is the projector of the observations, with the normalization condition $\mathbf{b}_{\cdot k}^\top \mathbf{b}_{\cdot k} = 1$, such that $\boldsymbol{F} = \boldsymbol{Z}\boldsymbol{B}$. We denote by $\lambda_{\cdot k}$ (or $2\mu_{\cdot k}$) the eigenvalues of the matrix $[GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top]$. Let the basis $\mathscr{B} := \{\mathbf{b}_{\cdot 1}, \ldots, \mathbf{b}_{\cdot h}\}$ with $h \leq K$ issued from the maximization of the overall Gini variability :

$$\max \mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k} \implies [GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top]\mathbf{b}_{\cdot k} = 2\mu_{\cdot k}\mathbf{b}_{\cdot k}, \ \forall k = 1, \ldots, K.$$

Indeed, from the Lagrangian,

$$L = \mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k} - \mu_{\cdot k}[1 - \mathbf{b}_{\cdot k}^\top \mathbf{b}_{\cdot k}],$$

because of the non-symmetry of $GC_\nu(\boldsymbol{Z})$, the eigenvalue equation is,

$$[GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top]\mathbf{b}_{\cdot k} = 2\mu_{\cdot k}\mathbf{b}_{\cdot k},$$

that is,

$$[GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top]\mathbf{b}_{\cdot k} = \lambda_{\cdot k}\mathbf{b}_{\cdot k}. \tag{3.10}$$

The new subspace $\{\mathbf{f}_{\cdot 1}, \ldots, \mathbf{f}_{\cdot h}\}$ such that $h \leq K$ is issued from the maximization of the Gini variability between the observations on each axis $\mathbf{f}_{\cdot k}$. Although the result of the generalized Gini PCA seems to be close to the classical PCA, some differences exist.

Proposition 3.4.1. *Let $\mathscr{B} = \{\mathbf{b}_{\cdot 1}, \ldots, \mathbf{b}_{\cdot h}\}$ with $h \leq K$ be the basis issued from the maximization of $\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k}$ for all $k = 1, \ldots, K$, then the following assertions hold :*
(i) $\max GMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = \mu_{\cdot k}$ *for all* $k = 1, \ldots, K$, *if and only if* $\boldsymbol{\mathrm{R}}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} = \boldsymbol{R}_{\mathbf{z}}^c\mathbf{b}_{\cdot k}$.
(ii) $\mathbf{b}_{\cdot k}\mathbf{b}_{\cdot h}^\top = 0$, *for all* $k \neq h$.
(iii) $\mathbf{b}_{\cdot k}\mathbf{b}_{\cdot k}^\top = 1$, *for all* $k = 1, \ldots, K$.

Démonstration. (i) Note that $GMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = -\frac{2\nu}{N(N-1)}(\mathbf{f}_{\cdot k} - \bar{\mathbf{f}})^\top \boldsymbol{\mathrm{R}}_{c,\mathbf{f}_{\cdot k}}^{\nu-1}$, where $\boldsymbol{\mathrm{R}}_{c,\mathbf{f}_{\cdot k}}^{\nu-1}$ is the kth column of the centered (decumulative) rank matrix $\boldsymbol{R}_{\mathbf{f}}^c$. Since $\mathbf{f}_{\cdot k} = \boldsymbol{Z}\mathbf{b}_{\cdot k}$ and $\bar{\mathbf{f}} = (\bar{\mathbf{f}}_{\cdot 1}, \ldots, \bar{\mathbf{f}}_{\cdot K}) = \mathbf{0}$ then : [9]

$$\begin{aligned}
\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k} &= -\frac{2\nu}{N(N-1)}\mathbf{b}_{\cdot k}^\top \boldsymbol{Z}^\top \boldsymbol{R}_{\mathbf{z}}^c\mathbf{b}_{\cdot k} \\
&= -\frac{2\nu}{N(N-1)}\mathbf{b}_{\cdot k}^\top \boldsymbol{Z}^\top \boldsymbol{\mathrm{R}}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} \quad (\text{by } \boldsymbol{\mathrm{R}}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} = \boldsymbol{R}_{\mathbf{z}}^c\mathbf{b}_{\cdot k}) \\
&= -\frac{2\nu}{N(N-1)}\mathbf{f}_{\cdot k}^\top \boldsymbol{\mathrm{R}}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} \\
&= GMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}).
\end{aligned} \tag{3.11}$$

Then, maximizing the multidimensional variability $\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k}$ yields from (3.10) :

$$\begin{aligned}
&\mathbf{b}_{\cdot k}^\top[GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top]\mathbf{b}_{\cdot k} = \mathbf{b}_{\cdot k}^\top\lambda_{\cdot k}\mathbf{b}_{\cdot k} \\
\iff\ &\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k} + \mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})^\top\mathbf{b}_{\cdot k} = \lambda_{\cdot k}.
\end{aligned}$$

9. We have :

$$\bar{\mathbf{f}}_{\cdot k} = 1/N\sum_{i=1}^N f_{ik} = 1/N\left[\sum_{i=1}^N \mathbf{z}_{i\cdot}^\top\mathbf{b}_{\cdot k}\right] = 1/N\left[\sum_{i=1}^N \mathbf{z}_{i\cdot}^\top\right]\mathbf{b}_{\cdot k} = \mathbf{0}.$$

Since $\mathbf{b}_{.k}^{\top} GC_{\nu}(\boldsymbol{Z})\mathbf{b}_{.k} = \mathbf{b}_{.k}^{\top} GC_{\nu}(\boldsymbol{Z})^{\top}\mathbf{b}_{.k}$, then

$$\mathbf{b}_{.k}^{\top} GC_{\nu}(\boldsymbol{Z})^{\top}\mathbf{b}_{.k} = \lambda_{.k}/2 = \mu_{.k},$$

and so $GMD_{\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.k}) = \lambda_{.k}$ for all $k = 1, \ldots, K$. The results (ii) and (iii) are straightforward. □

Discussion

Condition (i) shows that the maximization of the multidimensional variability (in the Gini sense) $\mathbf{b}_{.k}^{\top} GC_{\nu}(\boldsymbol{Z})\mathbf{b}_{.k}$ does not necessarily coincide with the maximization of the variability of the observations projected onto the new axis $\mathbf{f}_{.k}$ embodied by $GMD_{\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.k})$. Since in general, the rank of the observations on axis $\mathbf{f}_{.k}$ does not coincide with the projected ranks, that is,

$$\mathbf{R}_{c,\mathbf{f}_{.k}}^{\nu-1} \neq \boldsymbol{R}_{\mathbf{z}}^{c}\mathbf{b}_{.k},$$

then,

$$\max \mathbf{b}_{.k}^{\top} GC(\boldsymbol{Z})\mathbf{b}_{.k} \neq GMD_{\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.k}).$$

In other words, maximizing the quadratic form $\mathbf{b}_{.k}^{\top} GC(\boldsymbol{Z})_{\nu}\mathbf{b}_{.k}$ does not systematically maximize the overall Gini variability $GMD_{\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.k})$. However, it maximizes the following generalized Gini index :

$$\begin{aligned} GGMD_{\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.k}) &:= -\frac{2\nu}{N(N-1)}\mathbf{b}_{.k}^{\top}\boldsymbol{Z}^{\top}\boldsymbol{R}_{\mathbf{z}}^{c}\mathbf{b}_{.k} \\ &= -\frac{2\nu}{N(N-1)}\mathbf{f}_{.k}^{\top}\mathbf{b}_{.k}^{\top}(\boldsymbol{R}_{\mathbf{z}}^{c})^{\top}. \end{aligned}$$

In the literature on inequality indices, this kind of index is rather known as a generalized Gini index, because of the product between a variable $\mathbf{f}_{.k}$ and a function Ψ of its ranks, $\Psi(\mathbf{R}_{\mathbf{f}_k}) := \mathbf{b}_{.k}^{\top}(\boldsymbol{R}_{\mathbf{z}}^{c})^{\top}$, such that :

$$GGMD_{\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.k}) = -\frac{2\nu}{N(N-1)}\mathbf{f}_{.k}\ \Psi(\mathbf{R}_{\mathbf{f}_k}).$$

Yaari (1987) and subsequently Yaari (1988) proposes generalized Gini indices with a rank distortion function Ψ that describes the behavior of the decision maker (being either max-min or max-max). [10]

10. Strictly speaking Yaari (1987) and Yaari (1988) suggests probability distortion functions $\Psi : [0, 1] \to [0, 1]$, which does not necessarily coincide to our case.

It is noteworthy that this generalized Gini index of variability is very different from Banerjee (2010)'s multidimensional Gini index. The author proposes to extract the first eigenvector $\mathbf{e}_{\cdot 1}$ of $\boldsymbol{X}^\top \boldsymbol{X}$ and to project the data $\boldsymbol{X}$ such that $\mathbf{s} := \boldsymbol{X}\mathbf{e}_{\cdot 1}$ so that the multidimensional Gini index is $G(\mathbf{s}) = \mathbf{s}^\top \tilde{\Psi}(\mathbf{R}_s)$, with $\mathbf{R}_s$ the rank vector of $\mathbf{s}$ and with $\tilde{\Psi}$ a function that distorts the ranks. Banerjee (2010)'s index is derived from the matrix $\boldsymbol{X}^\top \boldsymbol{X}$. To be precise, the maximization of the variance-covariance matrix $\boldsymbol{X}^\top \boldsymbol{X}$ (based on the ℓ_2 metric) yields the projection of the data on the first component $\mathbf{f}_{\cdot 1}$, which is then employed in the multidimensional Gini index (based on the ℓ_1 metric). This approach is legitimated by the fact that $G(\mathbf{s})$ has some desirable properties linked with the Gini index. However, this Gini index deals with an information issued from the variance, because the vector $\mathbf{s}$ relies on the maximization of the variance of component $\mathbf{f}_{\cdot 1}$. Alternatively, it is possible to make use of the Gini variability, in a first stage, in order to project the data onto a new subspace, and in a second stage, to use the generalized Gini index of the projected data for the interpretations. In such as case, the Gini metric enables outliers to be attenuated. The employ of $G(\mathbf{s})$ as a result of the variance-covariance maximization may transform the data so that outlying observations would capture an important part of the information (variance) on the first component. This case occurs in the classical PCA. This fact will be proven in the next sections with Monte Carlo simulations. Let us before investigate the employ of the generalized Gini index $GGMD_\nu$.

Properties of $GGMD_\nu$

Since the Gini PCA relies on the generalized Gini index $GGMD_\nu$, let us explore its properties.

Proposition 3.4.2. *Let the eigenvalues of* $GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top$ *be such that* $\lambda_1 = \mu_1/2 \geq \cdots \geq \lambda_{\cdot K} = \mu_{\cdot K}/2$. *Then,*
(i) $GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot \ell}) = GGMD_\nu(\mathbf{f}_{\cdot \ell}, \mathbf{f}_{\cdot k}) = 0$, *for all* $\ell = 1, \ldots, K$, *if and only if,* $\lambda_{\cdot k} = 0$.
(ii) $\max_{k=1,\ldots,K} GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = \mu_1$.
(iii) $\min_{k=1,\ldots,K} GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = \mu_{\cdot K}$.

Démonstration. (i) The result comes from the rank-nullity theorem. From the eigenvalue Equation (3.10), we have :

$$\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k} = \lambda_{\cdot k}/2 = \mu_{\cdot k}.$$

Let f be the linear application issued from the matrix $GC_\nu(\boldsymbol{Z})$. Whenever $\lambda_{\cdot k} = 0$, two columns (or rows) of $GC_\nu(\boldsymbol{Z})$ are collinear, then the dimension of the image set of f is $\dim(f) = K - 1$. Hence, $\mathbf{f}_{\cdot k} = \mathbf{0}$. Since $\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})^\top \mathbf{b}_{\cdot k} = GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k})$ for all $k = 1, \ldots, K$, then for $\lambda_{\cdot k}$ we get :

$$\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})^\top \mathbf{b}_{\cdot k} = GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = \lambda_{\cdot k}/2 = \mu_{\cdot k} = 0.$$

On the other hand, since $\mathbf{f}_{\cdot k} = \mathbf{0}$, it follows that $GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot \ell}) = 0$ for all $\ell = 1, \ldots, K$. Also, if $\mathbf{f}_{\cdot k} = \mathbf{0}$ then the centered rank vector $\boldsymbol{\kappa}_{\mathbf{f}_{\cdot k}}^c = \mathbf{0}$, and so $GGMD_\nu(\mathbf{f}_{\cdot \ell}, \mathbf{f}_{\cdot k}) = 0$ for all $\ell = 1, \ldots, K$.
(ii) The proof comes from the Rayleigh-Ritz identity :

$$\lambda_{\max} := \max \frac{\mathbf{b}_1^\top [GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top] \mathbf{b}_1}{\mathbf{b}_1^\top \mathbf{b}_1} = \lambda_1.$$

Since $\mathbf{b}_1^\top GC_\nu(\boldsymbol{Z}) \mathbf{b}_{\cdot 1} = \lambda_1/2$ and because $\mathbf{b}_1^\top GC_\nu(\boldsymbol{Z}) \mathbf{b}_1 = GGMD_\nu(\mathbf{f}_1, \mathbf{f}_1)$, the result follows.
(iii) Again, the Rayleigh-Ritz identity yields :

$$\lambda_{\min} := \min \frac{\mathbf{b}_{\cdot K}^\top [GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top] \mathbf{b}_{\cdot K}}{\mathbf{b}_{\cdot K}^\top \mathbf{b}_{\cdot K}} = \lambda_K.$$

Then, $\mathbf{b}_{\cdot K}^\top GC_\nu(\boldsymbol{Z}) \mathbf{b}_{\cdot K} = GGMD_\nu(\mathbf{f}_{\cdot K}, \mathbf{f}_{\cdot K}) = \lambda_{\cdot K}/2$. □

The index $GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k})$ represents the variability of the observations projected onto component $\mathbf{f}_{\cdot k}$. When this variability is null, then the eigenvalue is null (i). In the same time, there is neither co-variability in the Gini sense between $\mathbf{f}_{\cdot k}$ and another axis $\mathbf{f}_{\cdot \ell}$, that is $GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot \ell}) = 0$.

In the Gaussian case, because the Gini correlation matrix is positive semi-definite, the eigenvalues are non-negative, then $GGMD$ is null whenever it reaches its minimum.

Proposition 3.4.3. *Let $Z \sim \mathcal{N}(0, 1)$ and let $\boldsymbol{X}$ represent identically distributed Gaussian random variables, with distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\rho})$ such that $\mathrm{Var}(X_k) = 1$ for all $k = 1, \ldots, K$ and let $\gamma_1, \ldots, \gamma_K$ be the eigenvalues of $\mathrm{Var}(\boldsymbol{X})$. Then the following assertions holds :*

(i) $\mathrm{Tr}[GC_\nu(\boldsymbol{X})] = \mathrm{Cov}(Z, \Phi(Z))\mathrm{Tr}[\mathrm{Var}(\boldsymbol{X})]$.
(ii) $\mu_k = \mathrm{Cov}(Z, \Phi(Z))\gamma_k$ *for all* $k = 1, \ldots, K$.
(iii) $|GC_\nu(\boldsymbol{X})| = \mathrm{Cov}^K(Z, \Phi(Z))|\mathrm{Var}(\boldsymbol{X})|$.
(iv) *For all $\nu > 1$:*

$$\frac{\mu_k}{\mathrm{Tr}[GC_\nu(\boldsymbol{X})]} = \frac{\gamma_k}{\mathrm{Tr}[\mathrm{Var}(\boldsymbol{X})]}, \ \forall k = 1, \ldots, K.$$

Démonstration. From Theorem 3.3.2 :

$$GC_\nu(\boldsymbol{X}) = \sigma^{-1}\text{Cov}(Z, \Phi(Z))Var(\boldsymbol{X}).$$

From Abramowitz & Stegun (1964) (Chapter 26), when $Z \sim \mathcal{N}(0,1)$,

$$\text{Cov}(Z, \Phi(Z)) = \frac{1}{2\sqrt{\pi}} \approx 0.2821.$$

Then the results follow directly. □

Point (iv) shows that the eigenvalues of the standard PCA are proportional to those issued from the generalized Gini PCA. Because each eigenvalue (in proportion of the trace) represents the variability (or the quantity of information) inherent to each axis, then both PCA techniques are equivalent when $\boldsymbol{X}$ is Gaussian :

$$\frac{\mu_k}{\text{Tr}[GC_\nu(\boldsymbol{X})]} = \frac{\gamma_k}{\text{Tr}[\text{Var}(\boldsymbol{X})]},\ \forall k = 1, \ldots, K\ ;\ \forall \nu > 1.$$

The $\mathbb{R}^N$-Euclidean space

In classical PCA, the duality between $\mathbb{R}^N$ and $\mathbb{R}^K$ enables the eigenvectors and eigenvalues of $\mathbb{R}^N$ to be deduced from those of $\mathbb{R}^K$ and conversely. This duality is not so obvious in the Gini PCA case. Indeed, in $\mathbb{R}^N$ the Gini variability between the observations would be measured by $GC_\nu(\widetilde{\boldsymbol{Z}}) := \frac{-2\nu}{N(N-1)}(\boldsymbol{R}^c_{\mathbf{z}})^\top \boldsymbol{Z}$, and subsequently the idea would be to derive the eigenvalue equation related to $\mathbb{R}^N$,

$$[GC_\nu(\widetilde{\boldsymbol{Z}}) + GC_\nu(\widetilde{\boldsymbol{Z}})^\top]\ \widetilde{\mathbf{b}}_{\cdot k} = \widetilde{\lambda}_{\cdot k}\widetilde{\mathbf{b}}_{\cdot k}.$$

The other option is to define a basis of $\mathbb{R}^N$ from a basis already available in $\mathbb{R}^K$. In particular, the set of principal components $\{\mathbf{f}_1, \ldots, \mathbf{f}_{\cdot k}\}$ provides by construction a set of normalized and orthogonal vectors. Let us rescale the vectors $\mathbf{f}_{\cdot k}$ such that :

$$\widetilde{\mathbf{f}}_{\cdot k} = \frac{\mathbf{f}_{\cdot k}}{GMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k})}.$$

Then, $\{\widetilde{\mathbf{f}}_1, \ldots, \widetilde{\mathbf{f}}_{\cdot k}\}$ constitutes an orthonormal basis of $\mathbb{R}^K$ in the Gini sense since $GMD_\nu(\widetilde{\mathbf{f}}_{\cdot k}, \widetilde{\mathbf{f}}_{\cdot k}) = 1$. This basis may be used as a projector of the

variables $\mathbf{z}_{\cdot k}$ onto $\mathbb{R}^N$. Let $\widetilde{\boldsymbol{F}}$ be the $N \times K$ matrix with $\widetilde{\mathbf{f}}_{\cdot k}$ in columns. The projection of the variables $\mathbf{z}_{\cdot k}$ in $\mathbb{R}^N$ is given by the following Gini correlation matrix :

$$\boldsymbol{V} := \frac{-2\nu}{N(N-1)}\, \widetilde{\boldsymbol{F}}^{\top} \boldsymbol{R}^{c}_{\mathbf{z}},$$

whereas it is given by $\frac{1}{N}\widetilde{\boldsymbol{F}}^{\top}\boldsymbol{Z}$ in the standard PCA, that is, the matrix of Pearson correlation coefficients between all $\widetilde{\mathbf{f}}_{\cdot k}$ and $\mathbf{z}_{\cdot \ell}$. The same interpretation is available in the Gini case. The matrix $\boldsymbol{V}$ is normalized in such a way that $\boldsymbol{V} \equiv [v_{k\ell}]$ are the G-correlations indices between $\widetilde{\mathbf{f}}_{\cdot k}$ and $\mathbf{z}_{\cdot \ell}$. This yields the ability to make easier the interpretation of the variables projected onto the new subspace.

3.5 Interpretations of the Gini PCA

The analysis of the projections of the observations and of the variables are necessary to provide accurate interpretations. Some criteria have to be designed in order to bring out, in the new subspace, the most significant observations and variables.

Observations

The absolute contribution of an observation i to the variability of a principal component $\mathbf{f}_{\cdot k}$ is :

$$ACT_{ik} = \frac{f_{ik}\ \Psi(R(f_{ik}))}{GGMD_{\nu}(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k})}.$$

The absolute contribution of each observation i to the generalized Gini Mean Difference of $\mathbf{f}_{\cdot k}$ (ACT_{ik}) is interpreted as a percentage of variability of $GGMD_{\nu}(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k})$, such that $\sum_{i=1}^{N} ACT_{ik} = 1$. This provides the most important observations i related to component $\mathbf{f}_{\cdot k}$ with respect to the information $GGMD_{\nu}(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k})$. On the other hand, instead of employing the Euclidean distance between one observation i and the component $\mathbf{f}_{\cdot k}$, the Manhattan distance is used. The relative contribution of an observation i to component $\mathbf{f}_{\cdot k}$ is then :

$$RCT_{ik} = \frac{|f_{ik}|}{\|\mathbf{f}_{i\cdot}\|_1}.$$

Remark that the gravity center of $\{\mathbf{f}_1, \ldots, \mathbf{f}_{.K}\}$ is $\mathbf{g} := (\bar{\mathbf{f}}_1, \ldots, \bar{\mathbf{f}}_{.K}) = \mathbf{0}$. The Manhattan distance between observation i and $\mathbf{g}$ is then $\sum_{k=1}^{K} |f_{ik} - 0|$, and so

$$RCT_{ik} = \frac{|f_{ik}|}{\|\mathbf{f}_{i.} - \mathbf{g}\|_1}.$$

The relative contribution RCT_{ik} may be interpreted rather as the contribution of dimension k to the overall distance between observation i and $\mathbf{g}$.

Variables

The most significant variables must be retained for the analysis and the interpretation of the data in the new subspace. It would be possible, in the same manner as in the observations case, to compute absolute and relative contributions from the Gini correlation matrix $\boldsymbol{V} \equiv [v_{k\ell}]$. Instead, it is possible to test directly for the significance of the elements $v_{k\ell}$ of $\boldsymbol{V}$ in order to capture the variables that significantly contribute to the Gini variability of components $\mathbf{f}_{.k}$. Let us denote $\widetilde{U}_{\ell k} := cov(\mathbf{f}_{.\ell}, \boldsymbol{R}^c_{\mathbf{z}_{.k}})$ with $\boldsymbol{R}^c_{\mathbf{z}_{.k}}$ the (decumulative) centered rank vector of $\mathbf{z}_{.k}$ raised to an exponent of $\nu - 1$ and $U_{.\ell} := cov(\mathbf{f}_{.\ell}, \boldsymbol{R}^c_{\mathbf{f}_{.\ell}})$. Those two Gini covariances yield the following U-statistics :

$$U_{\ell k} = \frac{\widetilde{U}_{\ell k}}{U_{.\ell}} = v_{k\ell}.$$

Let $U^0_{\ell k}$ be the expectation of $U_{\ell k}$, that is $U^0_{\ell k} := \mathbb{E}[U_{\ell k}]$. From Yitzhaki & Schechtman (2013), $U_{\ell k}$ is an unbiased and consistent estimator of $U^0_{\ell k}$. From Theorem 10.4 in Yitzhaki & Schechtman (2013), Chapter 10, we asymptotically get that $\sqrt{N}(U_{\ell k} - U^0_{\ell k}) \overset{a}{\sim} \mathcal{N}$. Then, it is possible to test for :

$$\left\| \begin{array}{l} H_0 : U^0_{\ell k} = 0 \\ H_1 : U^0_{\ell k} \neq 0. \end{array} \right.$$

Let $\hat{\sigma}^2_{\ell k}$ the Jackknife variance of $U_{\ell k}$, then it is possible to test for the null under the assumption $N \to \infty$ as follows : [11]

$$\frac{U_{\ell k}}{\hat{\sigma}_{\ell k}} \overset{a}{\sim} \mathcal{N}(0, 1).$$

11. As indicated by Yitzhaki (1991), the efficient Jackknife method may be used to find the variance of any U-statistics.

The usual PCA enables the variables to be analyzed in the circle of correlation, which outlines the correlations between the variables $\mathbf{z}_{.k}$ and the components $\mathbf{f}_{.\ell}$. In order to make a comparison with the usual PCA, let us rescale the U-statistics $U_{\ell k}$. Let $\mathbf{U}$ be the $K \times K$ matrix such that $\mathbf{U} \equiv [U_{\ell k}]$, and $\mathbf{u}_{.k}$ the k-th column of $\mathbf{U}$. Then, the absolute contribution of the variable $\mathbf{z}_{.k}$ to the component $\mathbf{f}_{.\ell}$ is :

$$\widetilde{ACT}_{k\ell} = \frac{U_{\ell k}}{\|\mathbf{u}_{.k}\|_2}.$$

The measure $\widetilde{ACT}_{k\ell}$ yields a graphical tool aiming at comparing the standard PCA with the Gini PCA. In the standard PCA, $\cos^2\theta$ (see Figure 3.2 below) provides the Pearson correlation coefficient between $\mathbf{f}_1$ and $\mathbf{z}_{.k}$. In the Gini PCA, $\cos^2\theta$ is the normalized Gini correlation coefficient $\widetilde{ACT}_{k1}$ thanks to the ℓ_2 norm.

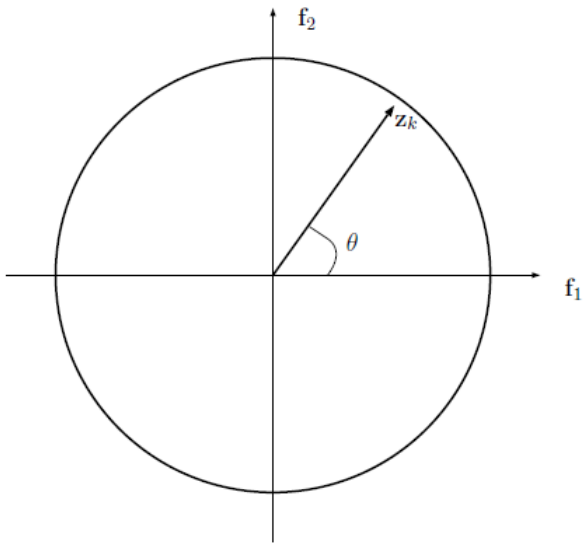


FIGURE 3.2 – Circle of correlation

It is worth mentioning that the circle of correlation does not provide the significance of the variables. This significance relies on the statistical test based on the U-statistics exposed before. Because $\widetilde{ACT}$ depends on the ℓ_2 metric, it is sensitive to outliers, and as such, the choice of the variables must rely on the test of $U^0_{\ell k}$ only.

3.6 Monte Carlo Simulations

In this Section, it is shown with the aid of Monte Carlo simulations that the usual PCA yields irrelevant results when outlying observations contaminate the data. To be precise, the absolute contributions computed in the standard PCA based on the variance may lead to select outlying observations on the first component in which there is the most important variability (a direct implication of the maximization of the variance). In consequence, the interpretation of the PCA may inflate the role of the first principal components. The Gini PCA dilutes the importance of the outliers to make the interpretations more robust and therefore more relevant.

Algorithm 4: Monte Carlo Simulation

Result: Robust Gini PCA with data contamination

$\theta = 1$ [θ is the value of the outlier] ;

repeat

  Generate a 4-variate normal distribution $\boldsymbol{X} \sim \mathcal{N}$, $N = 500$;

  Introduce outliers in 1 row of $\boldsymbol{X}$: $\boldsymbol{X}^o_{ji} := \theta \boldsymbol{X}_{ji}$ with $j = 1, \ldots, 4$ [for a random row localization];

  For each method (Variance and Gini), the ACT and RCT are computed for the axes 1 and 2 on the contaminated matrix $\boldsymbol{X}^o$;

until $\theta = 1000$ [increment of 1];

return Mean squared Errors of eigenvalues, ACT and RCT ;

The mean squared errors of the eigenvalues are computed as follows :

$$MSE_{\lambda_k} = \frac{\sum_{i=1}^{1,000} (\lambda_k^{oi} - \lambda_k)^2}{1,000},$$

where λ_k^{oi} is the eigenvalue computed with outlying observations in the sample. The MSE of ACT et RCT are computed in the same manner.

We first investigate the case where the variables are highly correlated in order to gauge the robustness of each technique (Gini for $\nu = 2, 4, 6$ and variance). The correlation matrix between the variables is given by :

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.8 & 0.9 & 0.7 \\ 0.8 & 1 & 0.8 & 0.75 \\ 0.9 & 0.8 & 1 & 0.6 \\ 0.7 & 0.75 & 0.6 & 1 \end{pmatrix}$$

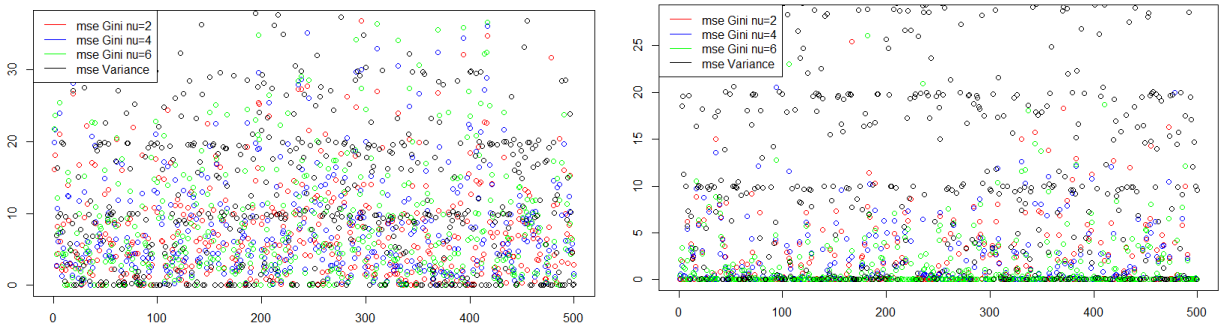
As can be seen in the matrix above, we can expect that all the information be gathered on the first axis because each pair of variables records an important linear correlation. The repartition of the information on each component, that is, each eigenvalue in percentage of the sum of the eigenvalues is the following.

Eigenvalues		Gini $\nu = 2$	Gini $\nu = 4$	Gini $\nu = 6$	Variance
Axis 1	eigenvalues (%)	81.65341	82.31098	82.28372	81.11458
	MSE	**12.313750**	**12.196975**	**12.221840**	**15.972710**
Axis 2	eigenvalues (%)	10.90079	10.47317	10.46846	11.35471
	MSE	**11.478541**	**11.204504**	**10.818344**	**10.688924**
Axis 3	eigenvalues (%)	5.062329	4.996538	5.088865	5.112817
	MSE	**2.605312**	**2.608647**	**2.799180**	**4.687323**
Axis 4	eigenvalues (%)	2.383476	2.219311	2.15896	2.417897
	MSE	**1.541453**	**1.826055**	**3.068596**	**2.295100**

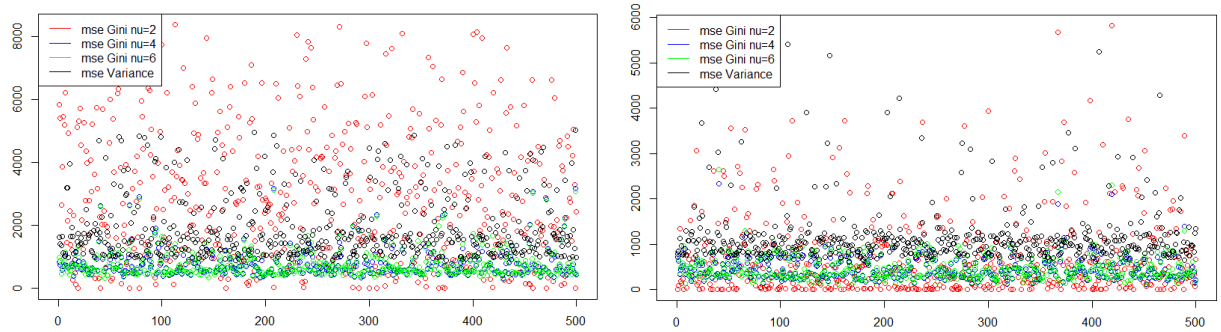
TABLE 3.1 – Eigenvalues and their MSE

The first axis captures around 82% of the variability of the overall sample (before contamination). Although each PCA method yields the same repartition of the information over the different components before the contamination of the data, it is possible to show that the classical PCA is not robust. For this purpose, let us analyze Figures 3.3a-3.3d below that depict the MSE of each observation with respect to the contamination process described in Algorithm 1 above.

On the first axis of Figure 3.3a, the absolute contribution of each observation (among 500 observations) is not stable because of the contamination of the data, however the Gini PCA performs better. The MSE of the ACTs measured during to the contamination process provides lower values for the Gini index compared with the variance. On the other hand, if we compute the standard deviation of all these MSEs over the two first axis, again the Gini methodology provides lower variations (see Table 3.2).



(a) ACT_1 (b) ACT_2



(c) RCT_1 (d) RCT_2

FIGURE 3.3 – ACT_1, ACT_2, RCT_1 and RCT_2

	Gini $\nu = 2$	Gini $\nu = 4$	Gini $\nu = 6$	Variance
Axis 1	6.08	6.62	7.41	**12.09**
Axis 2	4.07	5.12	13.37	2.98

TABLE 3.2 – Standard deviation of the MSE of the ACTs on the two first axis

Let us take now an example with less correlations between the variables in order to get a more equal repartition of the information on the first two axes.

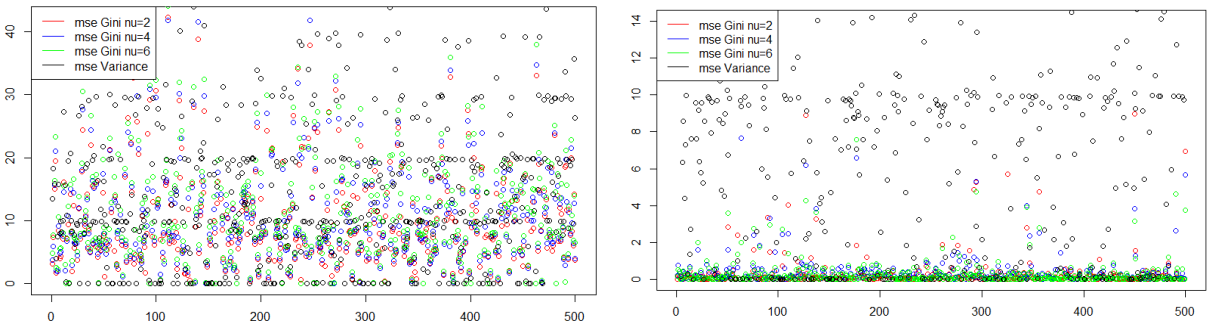
$$\rho = \begin{pmatrix} 1 & -0.5 & 0.25 & 0.5 \\ -0.5 & 1 & -0.9 & 1 \\ 0.25 & -0.9 & 1 & -0.25 \\ 0.5 & 0 & -0.25 & 1 \end{pmatrix}$$

The repartition of the information over the new axes (percentage of each eigenvalue) is given in Table 3.3. When the information is less concentrated on the first axis (55% on axis 1 and around 35% on axis 2), the MSE of the eigenvalues after contamination are much more important for the standard PCA compared with the Gini approach (2 to 3 times more important). Although the fourth axis reports an important MSE for the Gini method ($\nu = 6$), the eigenvalue percentage is not significant (1.56%).

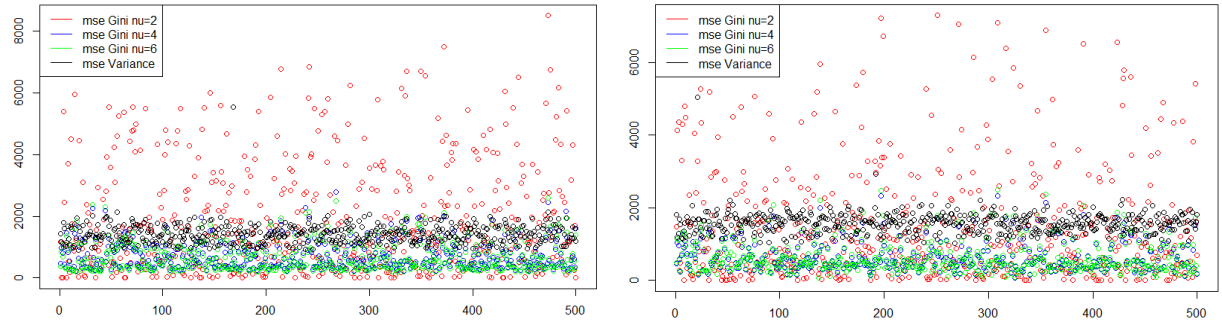
eigenvalues (%)		Gini $\nu = 2$	Gini $\nu = 4$	Gini $\nu = 6$	Variance
Axis 1	eigenvalues (%)	55.3774	55.15931	54.96172	55.08917
	MSE	**17.711023**	**14.968196**	**12.745760**	**38.929147**
Axis 2	eigenvalues (%)	35.8385	35.86216	35.8745	36.06118
	MSE	**14.012198**	**16.330350**	**18.929923**	**30.948674**
Axis 3	eigenvalues (%)	7.227274	7.345319	7.527222	7.329535
	MSE	**4.919686**	**4.897820**	**5.036241**	**6.814252**
Axis 4	eigenvalues (%)	1.556831	1.633214	1.636561	1.520114
	MSE	**1.149770**	**7.890184**	**14.047539**	**1.438904**

TABLE 3.3 – Eigenvalues and their MSE

Let us now have a look on the MSE of the absolute contributions of each observation ($N = 500$) for each PCA technique (3.4a-3.4b). We obtain the same kind of results, with less variability on the second axis. In Figures 3.4a-3.4b, it is apparent that the classical PCA based on the ℓ_2 norm exhibits much more ACT variability (black points). This means that the contamination of the data can lead to the interpretation of some observations as significant (important contribution to the variance of the axis) while they are not (and vice versa). On the other hand, the MSE of the RCTs after contamination of the data, Figures 3.4c-3.4d, are less spread out for the Gini technique for $\nu = 4$ and $\nu = 6$, however for $\nu = 2$ there is more variability of the MSE compared with the variance. This means that the distance from one observation to an axis may not be reliable (although the interpretation of the data rather depends on the ACTs).



(a) ACT_1 (b) ACT_2



(c) RCT_1 (d) RCT_2

FIGURE 3.4 – ACT_1, ACT_2, RCT_1 and RCT_2

The results of Figures 3.4a-3.4d can be synthesized by measuring the standard deviation of the MSE over the ACTs of the 500 observations along the two first axes.

	Gini $\nu = 2$	Gini $\nu = 4$	Gini $\nu = 6$	Variance
Axis 1	7.50	7.86	8.46	**11.92**
Axis 2	0.62	0.75	0.77	**11.24**

TABLE 3.4 – Standard deviation of the MSE of the ACT on the two first axes

As in the previous example of simulation, Table 3.4 indicates that the PCA based on the variance is less stable about the values of the ACTs that provide the most important observations of the sample. This may lead to irrelevant interpretations.

3.7 Application on cars data

We propose a simple application with the celebrated cars data (see the Appendix and our github). [12] The dataset is particularly interesting since there are highly correlated variables as can be seen in the Pearson correlation matrix given in Table 3.5.

	capacity x_1	**power** x_2	**speed** x_3	**weight** x_4	**width** x_5	**length** x_6
x_1	**1.000**	0.954	0.885	0.692	0.706	0.663
x_2	0.954	**1.000**	0.933	0.528	0.729	0.663
x_3	0.885	0.933	**1.000**	0.466	0.618	0.578
x_4	0.692	0.528	0.466	**1.000**	0.477	0.794
x_5	0.706	0.729	0.618	0.477	**1.000**	0.591
x_6	0.663	0.663	0.578	0.794	0.591	**1.000**

TABLE 3.5 – Correlation matrix

Also, the dataset is composed of some outlying observations (Figure 3.5) : Ferrari enzo (x_1, x_2, x_5), Bentley continental (x_2), Aston Martin (x_2), Land Rover discovery (x_5), Mercedes class S (x_5), Smart (x_5, x_6).

The overall information (variability) is partitioned over six components (Table 3.6).

eigenvalues in %	Gini $\nu = 2$	Gini $\nu = 4$	Gini $\nu = 6$	Variance
Axis 1	**80.35797**	**83.17172**	**84.84995**	**73.52112**
Axis 2	**12.0761**	**10.58655**	**9.715974**	**14.22349**
Axis 3	4.132136	2.987015	3.130199	**7.26106**
Axis 4	3.059399	2.612411	1.519626	3.93117
Axis 5	0.3332362	0.3125735	0.2696611	0.85727
Axis 6	0.04115858	-0.3297257	-0.5145944	0.20585
Sum	**100 %**	**100 %**	**100 %**	**100 %**

TABLE 3.6 – Eigenvalues (%)

Two axes may be chosen to analyze the data. As shown in the previous Section about the simulations, when the data are highly correlated such

12. An R markdown for Gini PCA is available : https://github.com/freakonometrics/GiniACP/
Data from Michel Tenenhaus's website (see also the Appendix) :
https://studies2.hec.fr/jahia/webdav/site/hec/shared/sites/tenenhaus/acces_anonyme/home/fichier_excel/auto_2004.xls

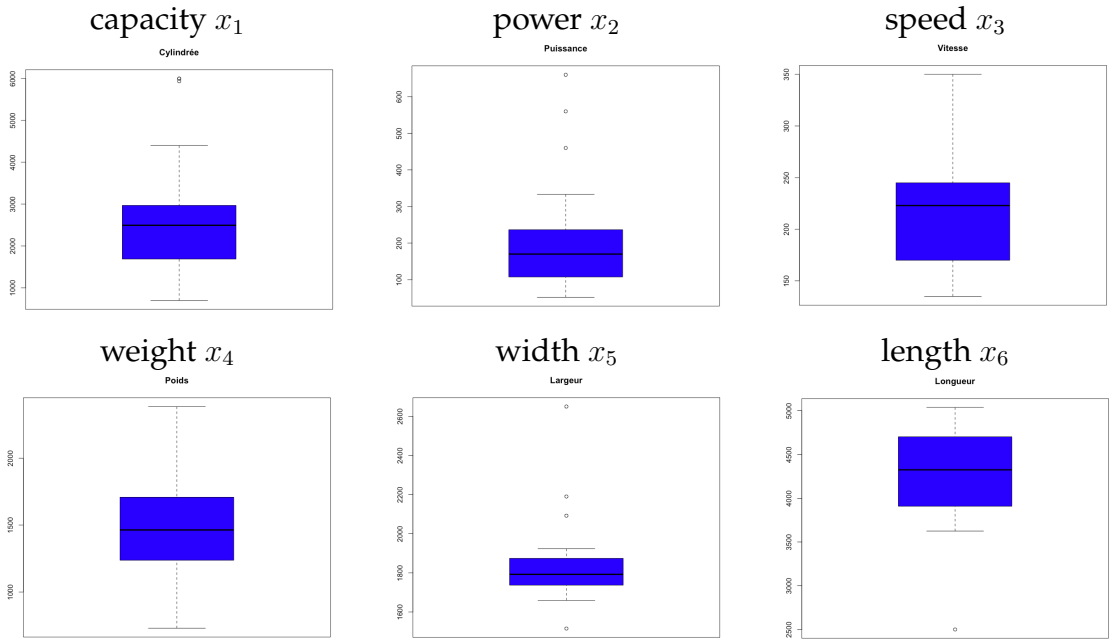
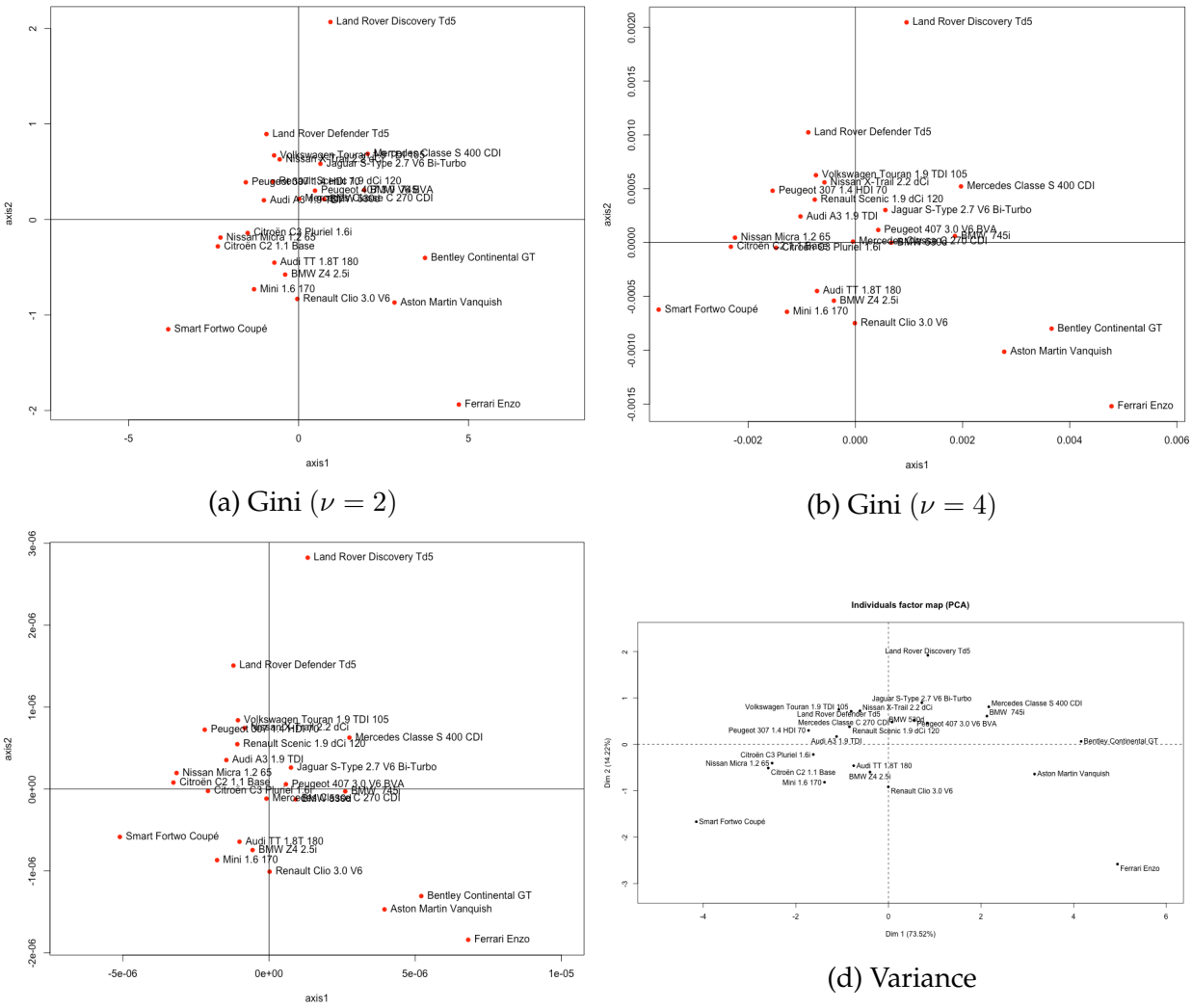


FIGURE 3.5 – Box plots

that two axes are sufficient to project the data, the Gini PCA and the standard PCA yield the same share of information on each axis. However, we can expect some differences for absolute contributions ACT and relative contributions RCT.

The projection of the data is depicted in Figure 3.6, for each method.

As depicted in Figure 3.6, the projection is very similar for each technique. The cars with extraordinary (or very low) abilities are in the same relative position in the four projections : Land Rover Discovery Td5 at the top, Ferrari Enzo at the bottom right, Smart Fortwo coupé at the bottom left. However, when we improve the coefficient of variability ν to look for what happens at the tails of the distributions (of the two axes), we see that more cars are distinguishable : Land Rover Defender, Audi TT, BMW Z4, Renault Clio 3.0 V6, Bentley Contiental GT. Consequently, contrary to the case $\nu = 2$ or the variance, the projections with $\nu = 4, 6$ allow one to find other important observations, which are not outlying observations that contribute to the overall amount of variability. For this purpose, let us first analyze the correlations between the variables and the new axes in order



(a) Gini ($\nu = 2$)

(b) Gini ($\nu = 4$)

(c) Gini ($\nu = 6$)

(d) Variance

FIGURE 3.6 – Projections of the cars

to interpret the results, see Tables 3.7 to 3.10.

Some slight differences appear between the Gini PCA and the classical one based on the variance. The theoretical Section 3.4 indicates that the Gini methodology for $\nu = 2$ is equivalent to the variance when the variables are Gaussian. On cars data, we observe this similarity. In each PCA, all variables are correlated with Axis 1 and weight with Axis 2. However,

when ν increases, the Gini methodology allows outlying observations to be diluted so that some variables may appear to be significant, whereas they are not in the variance case.

Gini ($\nu = 2$)		capacity	power	speed	weight	width	length
Axe 1	correlation	-0.974	-0.945	-0.872	0.760	-0.933	-0.823
	U**-stat**	**-56.416**	**-25.005**	**-10.055**	**-4.093**	**-24.837**	**-12.626**
Axe 2	correlation	-0.032	-0.241	-0.405	0.510	0.183	-0.379
	U**-stat**	**-0.112**	**-0.920**	**-1.576**	**2.897**	**0.526**	**1.666**

TABLE 3.7 – Correlations Axes / variables (significance 5%)

Gini ($\nu = 4$)		capacity	power	speed	weight	width	length
Axe 1	correlation	0.982	0.948	0.797	0.858	0.952	0.888
	U**-stat**	**8.990**	**8.758**	**4.805**	**9.657**	**8.517**	**8.182**
Axe 2	correlation	-0.021	0.207	0.516	-0.279	-0.147	-0.246
	U**-stat**	**-0.095**	**0.817**	**2.299**	**-1.773**	**-0.705**	**-1.200**

TABLE 3.8 – Correlations Axes / variables (significance 5%)

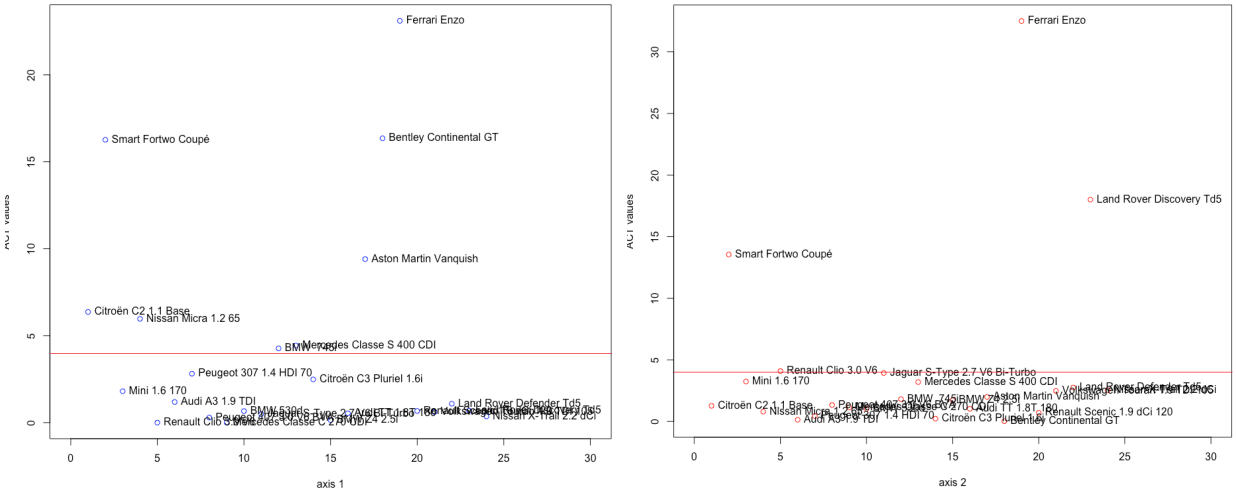
Gini ($\nu = 6$)		capacity	power	speed	weight	width	length
Axe 1	valeurs	-0.781	-0.759	-0.598	-0.730	-0.755	-0.701
	U**-stat**	**-4.036**	**-3.903**	**-3.137**	**-3.125**	**-3.882**	**-3.644**
Axe 2	valeurs	0.019	-0.170	-0.570	0.153	0.125	0.218
	U**-stat**	**0.089**	**-0.734**	**-1.914**	**0.734**	**0.569**	**0.906**

TABLE 3.9 – Correlations Axes / variables (significance 5%, 10%)

Variance		capacity	power	speed	weight	width	length
Axe 1	valeurs	0.962	0.923	0.886	0.756	0.801	0.795
	U**-stat**	**11.802**	**11.322**	**10.866**	**9.282**	**9.825**	**9.752**
Axe 2	valeurs	-0.126	-0.352	-0.338	0.575	-0.111	0.504
	U**-stat**	**-0.307**	**-0.855**	**-0.821**	**1.396**	**-0.269**	**1.223**

TABLE 3.10 – Correlations Axes / variables (significance 5%, 10%)

Tables 3.8 and 3.9 ($\nu = 4, 6$) show that Axis 2 is correlated to speed (not weight as in the variance PCA). In this respect the absolute contributions must describe the cars associated with speed on Axis 2. Indeed, the Land Rover discovery, a heavy weight car, is no more available on Axis 2 for the Gini PCA for $\nu = 2, 4, 6$ (Figures 3.8, 3.9, 3.10). Note that the red line in the Figures represents the mean share of the information on each axis, *i.e.* 100%/24 cars = 4.16% of information per car.

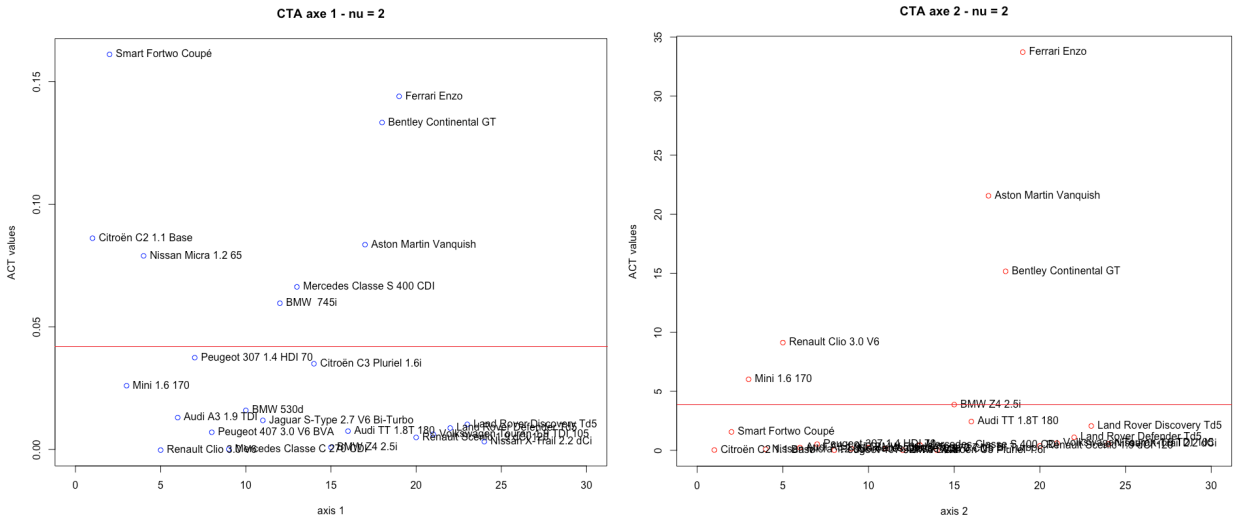


(a) Axis 1 (variance)

(b) Axis 2 (variance)

FIGURE 3.7 – Variance ACTs

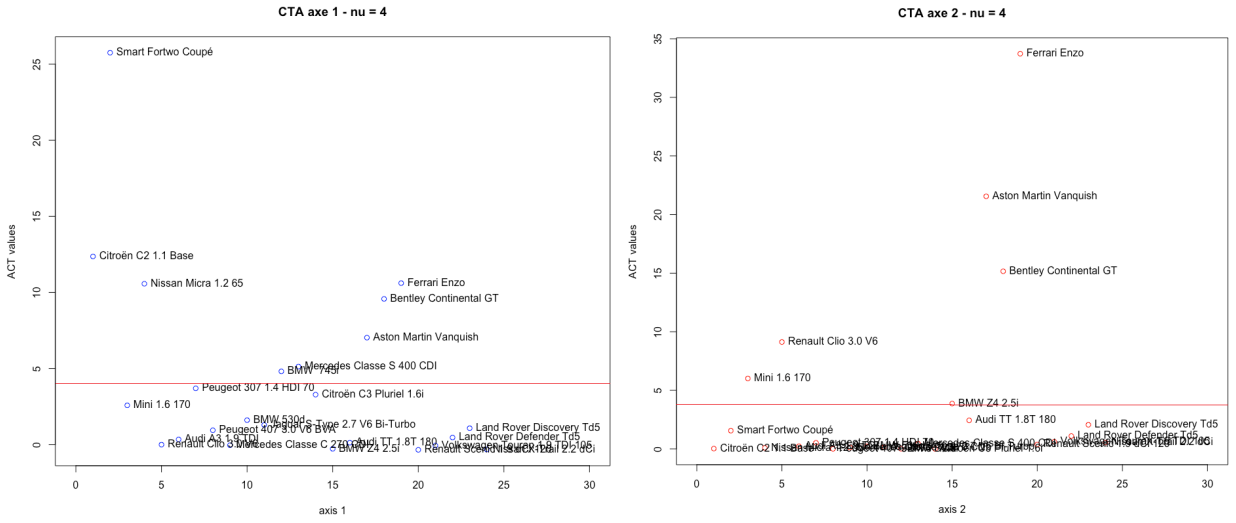
Finally, some cars are not correlated with axis 2 in the standard PCA, see Figures 3.8–3.10, while this is the case in the Gini PCA. Indeed some



(a) Axis 1 ($\nu = 2$)

(b) Axis 2 ($\nu = 2$)

FIGURE 3.8 – Gini ACTs ($\nu = 2$)

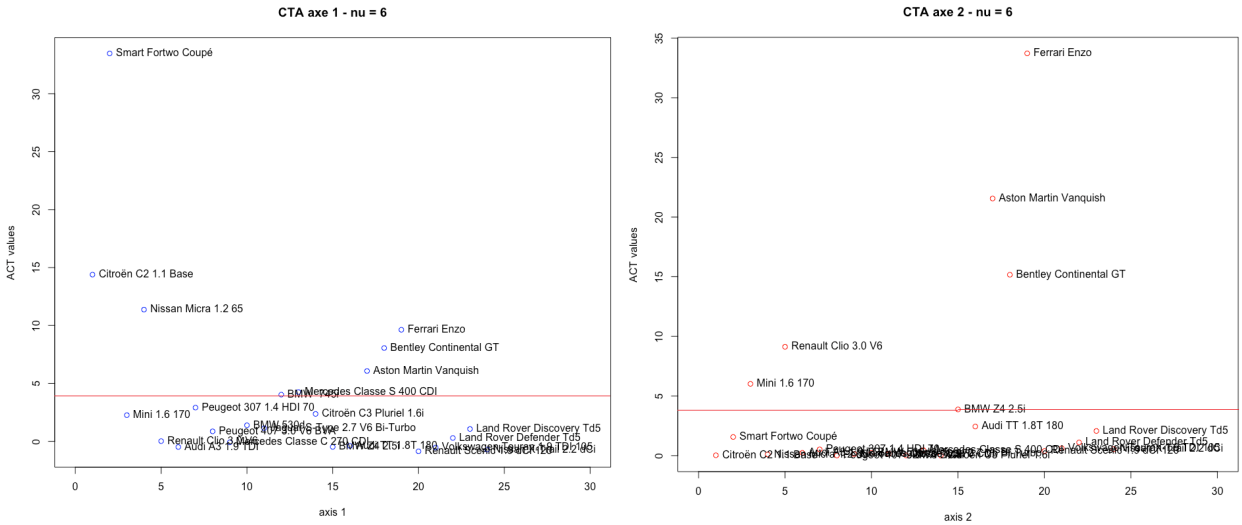


(a) Axis 1 ($\nu = 4$)

(b) Axis 2 ($\nu = 4$)

FIGURE 3.9 – Gini ACTs ($\nu = 4$)

cars are now associated with speed : Aston Martin, Bentley Continental GT, Renault Clio 3.0 V6 and Mini 1.6 170. This example of application shows



(a) Axis 1 ($\nu = 6$)

(b) Axis 2 ($\nu = 6$)

FIGURE 3.10 – Gini ACTs ($\nu = 6$)

that the use of the Gini metric robust to outliers may involve some serious changes in the interpretation of the results.

3.8 Conclusion

In this paper, it has been shown that the geometry of the Gini covariance operator allows one to perform Gini PCA, that is, a robust principal component analysis based on the ℓ_1 norm.

To be precise, the variance may be replaced by the Gini Mean Difference, which captures the variability of couples of variables based on the rank of the observations in order to attenuate the influence of the outliers. The Gini Mean Difference may be rather interpreted with the aid of the generalized Gini index $GGMD_\nu$ in the new subspace for a better understanding of the variability of the components, that is, $GGMD_\nu$ is both a rank-dependent measure of variability in Yaari (1987) sense and also an eigenvalue of the Gini correlation matrix.

Contrary to many approaches in multidimensional statistics in which the standard variance-covariance matrix is used to project the data onto

a new subspace before deriving multidimensional Gini indices (see e.g. Banerjee (2010)), we propose to employ the Gini correlation indices (see Yitzhaki & Schechtman (2013)). This provides the ability to interpret the results with the ℓ_1 norm and the use of U-statistics to measure the significance of the correlation between the new axes and the variables.

This research may open the way on data analysis based on Gini metrics in order to study multivariate correlations with categorical variables or discriminant analyses when outlying observations drastically affect the sample.

Appendix

cars	capacity x_1	power x_2	speed x_3	weight x_4	width x_5	length x_6
Citroën C2 1.1 Base	1124	61	158	932	1659	3666
Smart Fortwo Coupé	698	52	135	730	1515	2500
Mini 1.6 170	1598	170	218	1215	1690	3625
Nissan Micra 1.2 65	1240	65	154	965	1660	3715
Renault Clio 3.0 V6	2946	255	245	1400	1810	3812
Audi A3 1.9 TDI	1896	105	187	1295	1765	4203
Peugeot 307 1.4 HDI 70	1398	70	160	1179	1746	4202
Peugeot 407 3.0 V6 BVA	2946	211	229	1640	1811	4676
Mercedes Classe C 270 CDI	2685	170	230	1600	1728	4528
BMW 530d	2993	218	245	1595	1846	4841
Jaguar S-Type 2.7 V6 Bi-Turbo	2720	207	230	1722	1818	4905
BMW 745i	4398	333	250	1870	1902	5029
Mercedes Classe S 400 CDI	3966	260	250	1915	2092	5038
Citroën C3 Pluriel 1.6i	1587	110	185	1177	1700	3934
BMW Z4 2.5i	2494	192	235	1260	1781	4091
Audi TT 1.8T 180	1781	180	228	1280	1764	4041
Aston Martin Vanquish	5935	460	306	1835	1923	4665
Bentley Continental GT	5998	560	318	2385	1918	4804
Ferrari Enzo	5998	660	350	1365	2650	4700
Renault Scenic 1.9 dCi 120	1870	120	188	1430	1805	4259
Volkswagen Touran 1.9 TDI 105	1896	105	180	1498	1794	4391
Land Rover Defender Td5	2495	122	135	1695	1790	3883
Land Rover Discovery Td5	2495	138	157	2175	2190	4705
Nissan X-Trail 2.2 dCi	2184	136	180	1520	1765	4455

TABLE 3.11 – Cars data

BIBLIOGRAPHIE

Abramowitz, M. & I. Stegun. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series No. 55.

Anderson, T.W. (1963) Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, **34**, 122–148.

Baccini, A., P. Besse & A. de Falguerolles (1996), A L_1 norm PCA and a heuristic approach, in *Ordinal and Symbolic Data Analysis*, E Didday, Y. Lechevalier and O. Opitz (eds), Springer, 359–368.

Banerjee, A.K. (2010), A multidimensional Gini index, *Mathematical Social Sciences*, **60** : 87–93.

Brooks, J.P., Dulá, J.H. and E.L. Boone, A pure L1-norm principal component analysis, *Computational Statistics & Data Analysis*, **61**, 83-98.

Candes, E.J., Xiaodong Li, Yi Ma & John Wright. (2009) Robust Principal Component Analysis ?. arXiv :0912.3599.

Carcea, M. & R. Serfling (2015), A Gini autocovariance function for time series modeling. *Journal of Time Series Analysis* **36** : 817–38.

Charpentier A., Mussard S., and Ouraga T. (2020). *Principal Component Analysis : A Generalized Gini Approach*. European Journal of Operational Research, in under revew.

Dalton, H. 1920. The Measurement of the Inequality of Incomes. *The Economic Journal*, **30** :119, 348–361.

d'Aspremont, A., L. El Ghaoui, M.I. Jordan, & G. R. G. Lanckriet (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, **49** :3, 434–448.

Decancq, K. & M.-A. Lugo (2013), Weights in Multidimensional Indices of Well-Being : An Overview, *Econometric Reviews*, **32** :1, 7–34.

Ding, C., Zhou, D., He, X. & Zha, H. (2006). R_1-PCA : rotational invariant L1-norm principal component analysis for robust subspace factorization. *ICML '06 Proceedings of the 23rd international conference on Machine learning*, 281–288

Eckart, C. & G. Young (1936) The approximation of one matrix by another of lower rank. *Psychometrika*,**1**, 211–218.

Flury & Riedwyl (1988). Multivariate Statistics : A Practical Approach. Chapman & Hall

Furman, E. & R. Zitikis (2017), Beyond the Pearson Correlation : Heavy-Tailed Risks, Weighted Gini Correlations, and A Gini-Type Weighted Insurance Pricing Model, *ASTIN Bulletin : The Journal of the International Actuarial Association*, **47(03)** : 919-942.

Gajdos, T. and J. Weymark (2005), Multidimensional generalized Gini indices, *Economic Theory*, **26** :3, 471-496.

Gini, C. (1912), Variabilità e mutabilità, Memori di Metodologia Statistica, Vol. 1, Variabilità e Concentrazione. Libreria Eredi Virgilio Veschi, Rome, 211–382.

Giorgi, G.M. (2013), Back to the future : some considerations on Shlomo Yitzhaki and Edna Schechtman's book "'The Gini Methodology : A Primer on a Statistical Methodology"', *Metron*, **71(2)** : 189-195.

Gorban, A.N. , B. Kegl, D.C. Wunsch, & A. Zinovyev (Eds.) (2007) Principal Manifolds for Data Visualisation and Dimension Reduction. LNCSE 58, Springer Verlag.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417–441, 1933.

Korhonen, P. & Siljamäki, A. (1998). Ordinal principal component analysis theory and an application. *Computational Statistics & Data Analysis*, **26** :4, 411–424.

Kwak, N. (2008), Principal Component Analysis Based on L1-Norm Maximization, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30(9)**, 1672-1680.

Laurini, M.P. and A. Ohashi (2015), A noisy principal component analysis for forward rate curves, *European Journal of Operational Research*, **246(1)**, 140-153.

Li, C., Shaoa, Y.-H., & Deng, N.-Y. (2015) Robust L1-norm two-dimensional linear discriminant analysis, *Neural Networks*, **65** : 92-104.

List, C. (1999), Multidimensional Inequality Measurement : A Proposal, *Working paper*, Nuffield College.

Mackey, L. (2009) Deflation methods for sparse PCA. *Advances in Neural Information Processing Systems*, **21** : 1017–1024.

Mardia, K, Kent, J. & Bibby, J. (1979). Multivariate Analysis. Academic Press, London.

Olkin, Ingram, and Shlomo Yitzhaki. 1992. Gini regression analysis. *International Statistical Review* **60** : 185–96.

Pearson, K. (1901), On Lines and Planes of Closest Fit to System of Points in Space, *Philosophical Magazine*, **2** : 559–572.

Saad, Y. (1998). Projection and deflation methods for partial pole assignment in linear state feedback .*IEEE Trans. Automat. Contr.*, **33** : 290–297.

Schechtman, E. & S. Yitzhaki (1987), A Measure of Association Based on Gini's Mean Difference, *Communications in Statistics : A*, **16** : 207–231.

Schölkopf, B., Smola, A. and K. R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10(5), 1299-1319.

Schechtman, E. & S. Yitzhaki (2003), A family of correlation coefficients based on the extended Gini index, *Journal of Economic Inequality*, **1** :2, 129–146.

Shelef, A. (2016), A Gini-based unit root test. *Computational Statistics & Data Analysis*, **100** : 763–772.

Shelef, A., and E. Schechtman (2011), A Gini-based methodology for identifying and analyzing time series with non-normal innovations. *SSNR Electronic Journal* July : 1–26.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. *Journal of the Royal statistical society, series B*, **58** :1, 267–288.

Yaari, M.E. (1987), The Dual Theory of Choice Under Risk, *Econometrica*, **55** : 99–115.

Yaari, M.E. (1988), A Controversial Proposal Concerning Inequality Measurement, *Journal of Economic Theory*, **44** : 381–397.

Yitzhaki, S. (1991), Calculating Jackknife Variance Estimators for Parameters of the Gini Method. *Journal of Business and Economic Statistics*, **9** : 235–239.

Yitzhaki, S. (2003), Gini's Mean difference : a superior measure of variability for non-normal distributions, *Metron*, **LXI(2)** : 285-316.

Yitzhaki, S. & Olkin, I. (1991). Concentration indices and concentration curves. *Institute of Mathematical Statistics Lecture Notes*, **19** : 380–392.

Yitzhaki, S. & E. Schechtman (2013), *The Gini Methodology. A Primer on a Statistical Methodology*, Springer.

Zou, H., Hastie, T. & R. Tibshirani (2006), Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, **15** :2, 265-286.

CHAPITRE 4

SÉLECTION D'ACTIFS PAR ARBITRAGE, FRONTIÈRE EFFICIENTE ET RATIOS DE PERFORMANCE : UNE APPROCHE BASÉE SUR L'INDICE DE GINI

Sommaire

Résumé

Ce chapitre est une application à la finance des chapitres précédents. Il revisite le modèle multifactoriel de tarification des actifs, en particulier celui du modèle d'évaluation par arbitrage à facteurs latents (MEA-ACP) par une approche Gini. Le risque (ou variabilité) est mesuré par le Gini Mean Difference (GMD) et les facteurs latents sont obtenus suite à la maximisation de la matrice de corrélation au sens de Gini. Pour faire face à l'influence des outliers sur le calcul des sensibilités qui restent l'élément clé de la mise en place de ce modèle (MEA-ACP) et sur la fiabilité des facteurs latents, l'approche Gini est préconisée. Les facteurs latents et estimateurs par l'approche Gini sont donc utilisés pour la tarification des actifs par le MEA-ACP avec une approche Gini et pour calculer le ratio de performance de Treynor généralisé. Des simulations de Monte Carlo montrent la robustesse des primes de risque et du ratio de Treynor généralisé par l'approche Gini face à l'approche classique. Cette approche est illustrée par une application sur 8 actifs appartenant au marché financier français.

Mots-clés : Régression Gini, ACP, Robustesse, Arbitrage, Performance.

Abstract

A principal component analysis based on the generalized Gini correlation index is proposed (Gini PCA). The Gini PCA generalizes the This chapter is an application to finance of the previous chapters. It revisits the multifactor model of asset pricing, in particular that of the Latent Factor Arbitrage Valuation Model (APT-PCA) using a Gini approach. Risk (or variability) is measured by the Gini Mean Difference (GMD) and latent factors are obtained by maximizing the cGini correlation index. In order to cope with the influence of outliers on the calculation of sensitivities which remain the key element in the implementation of this model (APT-PCA) and on the reliability of latent factors, the Gini approach is recommended. The latent factors and estimators by the Gini approach are therefore used for the pricing of assets by the APT-PCA with a Gini approach and to calculate the generalized Treynor performance ratio. Monte Carlo simulations show the robustness of the risk premiums and the Treynor ratio generalized by

the Gini approach compared to the classical approach. This approach is illustrated by an application on 8 assets belonging to the French financial market.

Keywords : Gini Regression, PCA, Robustness, Arbitrage, Performance.

4.1 Introduction

La détection et l'estimation des risques systématiques liés aux rendements d'actifs demeurent des questions fondamentales en économie financière. Sur les marchés financiers, l'investisseur est confronté à un dilemme fondamental. Il doit faire le choix entre obtenir une rentabilité faible mais certaine et avoir une espérance de rentabilité élevée en se soumettant à une prise de risque importante. Le risque d'un titre ou d'un investissement peut être assimilé à la dispersion ou à la variabilité de sa rentabilité autour de sa valeur espérée. Dans la théorie financière moderne, le risque est généralement mesuré par la variance.

Le comportement des investisseurs est étudié afin de comprendre les mécanismes de formation des prix et des rentabilités sur les marchés financiers. De cette étude découlent les modèles d'équilibre fondés sur l'hypothèse d'efficience des marchés financiers, selon laquelle le prix des actifs financiers reflète spontanément et pleinement toute l'information relative à ce titre. A partir des travaux fondateurs de Markowitz (1952 & 1959), Sharpe (1964), Lintner (1965) et Mossin (1966), le modèle d'évaluation des actifs financiers (MEDAF ou CAPM) s'est développé. Ce modèle stipule que les rentabilités des titres sont des fonctions linéaires d'un seul facteur de risque commun – le rendement du marché– avec des hypothèses assez restrictives quant aux préférences des investisseurs. Comme la plupart des modèles scientifiques, plusieurs critiques sont adressées à ce modèle dont la plus importante est celle de Fama et French (1992). Les travaux de Fama et French concluent que seul le rendement du marché ne peut correctement expliquer le rendement moyen à long terme d'un actif risqué. Pour faire face à cette limite majeure du MEDAF, plusieurs modèles multifactoriels sont proposés par des auteurs tels que Ross (1976) avec le modèle d'évaluation par arbitrage (MEA ou APT [1]), Fama et French (1992), Carhart (1997), Mbairadjim Moussa et Sadefo Kamdem (2016). Le premier facteur pouvant être le portefeuille de marché comme dans la MEDAF. Ce portefeuille de marché est déterminé à partir des actifs risqués et de l'actif sans risque. Le portefeuille de marché (PM) sera le portefeuille de tangence avec la demi-droite qui lie l'actif sans risque et la frontière efficiente formée par l'ensemble des titres risqués. Dans la pratique, trois types de modèles

1. Arbitrage Pricing Theory.

multifactoriels ont été proposés.

Le premier type est le modèle à facteurs caractéristiques des entreprises. Ce modèle considère que les caractéristiques d'une entreprise, sont des éléments importants qui permettent d'expliquer les différences de rentabilité entre les titres financiers. Les facteurs caractéristiques peuvent être classés en deux grands groupes : les facteurs caractéristiques fondamentaux et les facteurs caractéristiques de marché. Les premiers sont associés aux caractéristiques de l'activité de l'entreprise (endettement, taille, etc.) et les seconds sont associés au comportement du cours boursier de l'entreprise (volume de transaction, momentum [2], etc.). Les valeurs caractéristiques ont de manière générale une influence positive sur certaines périodes et négatives sur d'autres.

Le second type est le modèle à facteurs macroéconomiques. Il s'agit de trouver des facteurs économiques, sources de risque communs à toutes les entreprises du marché financier concerné. Deux types de facteurs sont donc mis en avant. D'une part des facteurs économiques qui affectent l'activité de toutes les entreprises, et d'autre part, des facteurs économiques qui affectent le comportement des investisseurs dans ces entreprises et donc des règles de valorisation des entreprises sur les marchés financiers. Le troisième type est le modèle à facteurs statistiques. En utilisant une base de données sur l'historique des cours de clôture boursiers, les rentabilités boursières d'un nombre important de titres sont calculées. A partir de ces rentabilités, il est possible d'extraire un petit nombre de facteurs "'implicites"' grâce à des techniques statistiques d'analyse des données connues comme l'analyse en composantes principales (ACP). Ces facteurs donnent la meilleure explication des rentabilités antérieures des titres. En analyse de données, pour analyser la matrice qui résulte de la collecte de donnés, le critère de variance multidimensionnelle est largement utilisé. Le problème est que la variance capture une notion très précise de la dispersion qui ne correspond pas toujours aux propriétés que satisfont d'autres indices de variabilité ou d'inégalité tel que l'indice de Gini.

Il est proposé dans ce Chapitre, d'utiliser l'ACP-Gini développée dans les chapitres précédents afin de trouver les facteurs du modèle à facteurs statistiques. Plus récemment, Banerjee (2010) montre qu'il est possible de

2. Le taux d'accélération du prix d'un titre ou de son volume.

construire des indices de Gini multidimensionnels en prenant en compte les vecteurs propres de la matrice étudiée, néanmoins ces vecteurs propres sont issus da la matrice de variances-covariances. En recourant à une ACP de type Gini plusieurs questions se posent :

Comment trouver la valeur fondamentale des titres à l'équilibre ? Quelle mesure de risque et modèle d'estimation conviendraient pour estimer ces modèles d'évalution des actifs financiers ?

Dans ce Chapitre, il est appliqué une approche à facteurs statistiques. Nous appliquons une ACP-Gini sur un grand nombre de titres risqués (actions) sélectionnés afin de déterminer un petit nombre de facteurs "implicites". Aussi vrai qu'il est difficile de mesurer correctement le portefeuille de marché dans le cas du MEDAF, la structure à tester quant à elle reste bien définie. Ce qui n'est pas le cas pour les modèles à plusieurs facteurs comme l'APT.

Ce chapitre s'articule autour de deux grandes parties. La première partie est consacrée à l'évaluation des risques financiers. Il s'agit de présenter d'abord la gestion de portefeuilles à travers l'approche de la diversification, ensuite les modèles d'équilibre des actifs financiers (MEDAF et MEA) et ensuite présenter le ratio de performance généralisé de Treynor qui s'appuit sur le MEA. La deuxième partie envisage la conception des modèles d'évaluation des risques financiers de la première partie grâce aux outils de la méthodologie Gini par une étude de cas. Ce travail est finalement conclu dans une section où nous tirons différents enseignements de l'étude de cas.

4.2 Évaluation des risques financiers

L'évaluation des actifs est une estimation de la valeur théorique d'un actif. Les modèles à facteurs nous permettent de préciser la mesure de risque et d'évaluer la prime de risque liée à la détention d'un actif ou d'un portefeuille. En plus du risque à évaluer et à quantifier, il est nécessaire et indispensable de mesurer la performance de l'investissement réalisé. Cette section comporte quatre parties. La première partie est consacrée à la gestion de portefeuille qui consiste à sélectionner des portefeuilles "'efficients"' avec une approche Gini du risque. La deuxième partie est le *modèle d'évaluation des actifs financiers* (MEDAF) qui stipule que seul le risque de marché [3] influence l'espérance de rendement d'un actif ou d'un portefeuille. La troisième partie traite du *modèle d'évaluation par arbitrage* (MEA) qui est une théorie du prix du risque et un modèle multifactoriel qui s'appuie sur le principe d'arbitrage [4]. La quatrième partie introduit le ratio de performance généralisé de Treynor qui est défini comme l'excès de rendement d'un portefeuille par unité de moyenne pondérée de risque systématique, normalisé par la moyenne pondérée de la prime de risque systématique d'un benchmark [5].

Diversification, frontière efficiente et portefeuille de marché

Toute politique de placement vise à chercher une rentabilité élevée tout en contrôlant le risque de ce placement. Ainsi, à chaque placement est associé le couple rentabilité/risque. Afin de réduire le risque global pris par un investisseur sans pour autant sacrifier sa rentabilité à long terme, une technique ou stratégie de gestion de portefeuille appelée *diversification* est utilisée. La diversification consiste à mélanger à l'intérieur d'un portefeuille plusieurs actifs ou plusieurs classes d'actifs (actions, obligations,

3. Le risque de marché est matérialisé par l'évolution tendancielle du cours de la valeur par rapport à la moyenne du marché.

4. Le modèle MEA est basé sur la loi du prix unique. Cette loi stipule que dans un marché performant ou efficient, les actifs ou les portefeuilles ayant le même niveau de risque devraient s'échanger au même prix.

5. D'après le lexique financier de "'Vernimnen.net"', un benchmark est un très bon niveau de performance atteint par des acteurs dans un secteur et qui sert de référence aux acteurs moins performants pour essayer d'amener leurs propres performances au niveau de celles du benchmark.

immobiliers, etc.). Un portefeuille bien diversifié donne des avantages à l'investisseur. En cas de crise économique ou simplement en cas de retournement de marché, l'investisseur bénéficiera d'un effet de compensation entre certains actifs présents dans son portefeuille qui viennent atténuer la volatilité à la baisse. En l'absence de crise, cette stratégie permet de réaliser des rentabilités globales plus régulières.

En 1959, Markowitz introduit la théorie de la frontière efficiente (ou théorie moyenne-variance) qui consiste à sélectionner des portefeuilles optimaux en prenant en compte à la fois la rentabilité espérée (moyenne) et l'incertitude sur cette rentabilité (le risque). Le risque dans la théorie de Markowitz tout comme la théorie moderne du portefeuille est mesuré par la variance. Dans ce chapitre le risque est mesuré par la différence moyenne de Gini ou Gini's Mean Difference (GMD). Le GMD mesure l'écart absolu de rendement espéré entre deux périodes prises au hasard sur la période d'étude. Le portefeuille de marché, quant à lui, se situe sur la frontière efficiente et contient l'ensemble des titres risqués présents sur le marché en proportion de leurs capitalisations boursières.

Résultats théoriques

En utilisant la matrice Gini-Cogini $\boldsymbol{G} \equiv [GMD(\mathbf{x}_l, \mathbf{x}_k)]$, les étapes de construction de la frontière efficiente et du portefeuille de marché sont :

1. Calculer les poids du portefeuille à risque minimal global $\mathbf{p}_m^*$ avec $\mathbf{p}_m = (p_{m1}, p_{m2}, \cdots, p_{mn})'$;
2. Calculer les poids du portefeuille efficient $\mathbf{p}$ pour une espérance de rendement α égale au maximum des espérances de rendement des actifs constituants le portefeuille $\alpha = \max\{\mu_1, \mu_2, \cdots, \mu_n\}$;
3. Utiliser ces deux portefeuilles appartenants à la frontière efficiente pour former de nouveaux portefeuilles efficients $\mathbf{z}$;
4. Créer un compteur $\theta = 1, 0.9, \cdots, -1$ qui permet de construire les nouveaux poids $\mathbf{z}$ des portefeuilles de la frontière efficiente tel que $\mathbf{z} = \theta\, \mathbf{p}_m^* + (1 - \theta)\, \mathbf{p}$;
5. Ajouter un actif sans risque à l'univers d'actifs risqués puis déterminer le portefeuille qui minimise le GMD pour un excès de rendement cible.

Theorem 4.2.1. *Soit $\boldsymbol{G}_{add}$, la somme de la matrice de variabilité au sens de Gini et de sa transposée $\boldsymbol{G}_{add} = \boldsymbol{G} + \boldsymbol{G}'$. Pour $\boldsymbol{G}$ inversible, le vecteur $\boldsymbol{p}_m^* = \frac{\boldsymbol{G}_{add}^{-1}\mathbb{1}}{\mathbb{1}'\boldsymbol{G}_{add}^{-1}\mathbb{1}}$ est le vecteur des poids du portefeuille à risque minimal. Autrement dit, $\boldsymbol{p}_m^*$ est la solution du problème de minimisation suivant :*

$$\begin{aligned} \min \Gamma^P &= \min \boldsymbol{p}_m' \, \boldsymbol{G} \, \boldsymbol{p}_m \\ \textit{s.c. } \boldsymbol{p}_m' \mathbb{1} &= 1. \end{aligned}$$

Démonstration. Nous savons que le Gini du portefeuille est :

$$\Gamma^P = \mathbf{p}_m' \, \boldsymbol{G} \, \mathbf{p}_m,$$

avec $\boldsymbol{G}$ la matrice Gini-Cogini et $\mathbf{p}_m = (p_{m1}, p_{m2}, \cdots, p_{mn})'$. Le lagrangien s'écrit :

$$L(\mathbf{p}_m, \lambda) = \mathbf{p}_m' \boldsymbol{G} \, \mathbf{p}_m + \lambda(1 - \mathbf{p}_m' \mathbb{1}), \tag{4.1}$$

où λ est une constante positive. Les conditions de premier ordre (CPO) sont :

$$\frac{\partial \, L(\mathbf{p}_m, \lambda)}{\partial \, \mathbf{p}_m} = 0 \iff (\boldsymbol{G} + \boldsymbol{G}')\mathbf{p}_m - \lambda \mathbb{1} = 0.$$

Posons $\boldsymbol{G} + \boldsymbol{G}' = \boldsymbol{G}_{add}$, alors :

$$\boldsymbol{G}_{add} \, \mathbf{p}_m - \lambda \mathbb{1} = 0 \tag{4.2}$$

$$\frac{\partial \, L(\mathbf{p}_m, \lambda)}{\partial \, \lambda} = 0 \iff 1 - \mathbf{p}_m' \mathbb{1} = 0. \tag{4.3}$$

De l'équation (2), $\mathbf{p}_m = \lambda \, \boldsymbol{G}_{add}^{-1} \, \mathbb{1}$. En multipliant chaque membre de cette équation par $\mathbb{1}'$:

$$\begin{aligned} \mathbb{1}'\mathbf{p}_m &= \lambda \, \mathbb{1}' \, \boldsymbol{G}_{add}^{-1} \, \mathbb{1} \\ 1 &= \lambda \, \mathbb{1}' \, \boldsymbol{G}_{add}^{-1} \, \mathbb{1} \\ \lambda &= \frac{1}{\mathbb{1}' \, \boldsymbol{G}_{add}^{-1} \, \mathbb{1}} \end{aligned} \tag{4.4}$$

Nous avons $\mathbf{p}_m = \lambda \, \boldsymbol{G}_{add}^{-1} \, \mathbb{1}$. Donc en utilisant l'expression de λ :

$$\begin{aligned}
\mathbf{p}_m &= \lambda\, \boldsymbol{G}_{add}^{-1}\, \mathbb{1} \\
\mathbf{p}_m &= \frac{1}{\mathbb{1}'\, \boldsymbol{G}_{add}^{-1}\, \mathbb{1}}\, \boldsymbol{G}_{add}^{-1}\, \mathbb{1} \\
\mathbf{p}_m^* &= \frac{\boldsymbol{G}_{add}^{-1}\, \mathbb{1}}{\mathbb{1}'\, \boldsymbol{G}_{add}^{-1}\, \mathbb{1}} \qquad (4.5)
\end{aligned}$$

□

Le rendement espéré du portefeuille à risque minimal est alors :

$$\mathbf{p}_m'\boldsymbol{\mu} = (p_{m1}, p_{m2}, \cdots, p_{mn}) \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_n \end{pmatrix} = \alpha_m$$

Theorem 4.2.2. *Soit :*

$$\boldsymbol{z}_p = \underbrace{\begin{pmatrix} \boldsymbol{G}_{add} & \boldsymbol{\mu} & \mathbb{1} \\ \boldsymbol{\mu}' & 0 & 0 \\ \mathbb{1}' & 0 & 0 \end{pmatrix}^{-1}}_{\boldsymbol{A}_p} \underbrace{\begin{pmatrix} 0 \\ \alpha \\ 1 \end{pmatrix}}_{\boldsymbol{b}_0}$$

Le vecteur $\boldsymbol{P} = (p_1, \ldots, p_n)$, issu des n premières valeurs de $\mathbf{z}_p$, est pour un rendement espéré du portefeuille $\alpha > 0$ donné et pour $\boldsymbol{G}$ inversible, la solution du problème de minimisation suivant :

$$\begin{aligned}
& & \min \Gamma^P &= \min \boldsymbol{p}'\, \boldsymbol{G}\, \boldsymbol{p} \\
& s.c. & \boldsymbol{p}'\boldsymbol{\mu} &= \alpha \\
& & \boldsymbol{p}'\mathbb{1} &= 1
\end{aligned}$$

avec $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)$.

Démonstration. Le lagrangien s'écrit :

$$L(\mathbf{p}, \lambda_1, \lambda_2) = \mathbf{p}'\, \boldsymbol{G}\, \mathbf{p} + \lambda_1(\mathbf{p}'\mathbb{1} - 1) + \lambda_2(\mathbf{p}'\mu - \alpha) \qquad (4.6)$$

où λ_1 et λ_2 sont des constantes positives. Les CPO sont :

$$\frac{\partial L(\mathbf{p}, \lambda_1, \lambda_2)}{\partial \mathbf{p}} = 0 \iff [\boldsymbol{G} + \boldsymbol{G}']\mathbf{p} + \lambda_1 \mathbb{1} + \lambda_2 \boldsymbol{\mu} = 0$$

Posons $\boldsymbol{G} + \boldsymbol{G}' = \boldsymbol{G}_{add}$, alors : $\boldsymbol{G}_{add}\mathbf{p} + \lambda_1 \mathbb{1} + \lambda_2 \boldsymbol{\mu} = 0$.

$$\frac{\partial L(\mathbf{p}, \lambda_1, \lambda_2)}{\lambda_1} = 0 \iff \mathbf{p}'\mathbb{1} - 1 = 0 \tag{4.7}$$

$$\frac{\partial L(\mathbf{p}, \lambda_1, \lambda_2)}{\lambda_2} = 0 \iff \mathbf{p}'\boldsymbol{\mu} - \alpha = 0 \tag{4.8}$$

Ces CPO consistent à résoudre $n + 2$ équations linéaires à $n + 2$ inconnus $(p_1, p_2, \cdots, p_n, \lambda_1, \lambda_2)$. Réécrivons les CPO sous forme matricielle :

$$\underbrace{\begin{pmatrix} \boldsymbol{G}_{add} & \boldsymbol{\mu} & \mathbb{1} \\ \boldsymbol{\mu}' & 0 & 0 \\ \mathbb{1}' & 0 & 0 \end{pmatrix}}_{\mathbf{A}_p} \underbrace{\begin{pmatrix} \mathbf{p} \\ \lambda_1 \\ \lambda_2 \end{pmatrix}}_{\mathbf{z}_p} = \underbrace{\begin{pmatrix} 0 \\ \alpha \\ 1 \end{pmatrix}}_{\mathbf{b}_0} \tag{4.9}$$

ou,

$$\mathbf{A}_p \, \mathbf{z}_p = \mathbf{b}_0$$

La solution pour $\mathbf{z}_p$ est alors :

$$\mathbf{z}_p = \mathbf{A}_p^{-1} \, \mathbf{b}_0 \tag{4.10}$$

Les n premiers éléments de $\mathbf{z}_p$ sont les poids $\mathbf{p} = (p_1, p_2, \cdots, p_n)'$ du portefeuille efficient avec une espérance de rendement $\mathbf{p}'\mu = \alpha$. □

En plus des actifs risqués, pour le portefeuille tangent (ou portefeuille de marché) il existe un actif sans risque rf dans l'univers des titres. Il existe un portefeuille efficient qui se caractérise par la présence d'un actif additionnel sans risque rf. L'ancienne frontière efficiente devient une demie droite qui part de l'actif sans risque et passe par le portefeuille tangent. Ce portefeuille appartient à la famille des portefeuilles se trouvant sur la droite des marchés de capitaux ou Capital Market Line (CML) en anglais. Dans l'espace moyenne-GMD, la CML est la droite formée par l'ensemble des portefeuilles composés de l'actif sans risque et du portefeuille tangent. Considérons des actifs risqués avec un vecteur de rendement $\mathbf{r}$ et un actif sans risque qui rapporte un rendement r_f. Soit $\mathbf{p}$ le vecteur poids des

actifs risqués et p_f le poids de l'actif sans risque tel que $\mathbf{x}'\mathbb{1} + x_f = 1$. Le rendement du portefeuille est alors :

$$\begin{aligned} R_p^p &= \mathbf{p}'\mathbf{r} + p_f r_f \\ &= \mathbf{p}'\mathbf{r} + (1 - \mathbf{p}'\mathbb{1}) r_f \\ &= r_f + \mathbf{p}'(\mathbf{r} - r_f \mathbb{1}) \end{aligned}$$

L'excès de rendement du portefeuille est :

$$R_p^p - r_f = \mathbf{p}'(\mathbf{r} - r_f \mathbb{1}) \tag{4.11}$$

La prime de risque (excès de l'espérance de rendement) et le GMD du portefeuille sont :

$$\begin{aligned} \mu_p^p - r_f &= \mathbf{p}'(\boldsymbol{\mu} - r_f \mathbb{1}) \text{ avec } \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n) \\ \Gamma^P &= \mathbf{p}' \, \boldsymbol{G} \, \mathbf{p} \end{aligned} \tag{4.12}$$

Afin de simplifier les écritures, définissons :

$$\begin{aligned} \bar{\mathbf{r}} &= \mathbf{r} - r_f \mathbb{1} \\ \bar{\boldsymbol{\mu}} &= \boldsymbol{\mu} - r_f \mathbb{1} \\ \widetilde{R}_p^p &= R_p^p - r_f \\ \widetilde{\mu}_p^p &= \mu_p^p - r_f \end{aligned}$$

Les équations 4.25 et 4.26 peuvent être réécrites comme :

$$\widetilde{R}_p^p = \mathbf{p}'\bar{\mathbf{r}} \tag{4.13}$$

$$\widetilde{\mu}_p^p = \mathbf{p}'\bar{\boldsymbol{\mu}} \tag{4.14}$$

Theorem 4.2.3. *Soit $\widetilde{\mu}_p^0 = \mu_p^0 - r_f$ un excès de rendement cible. Le vecteur des poids $\boldsymbol{p} = \widetilde{\mu}_p^0 \frac{\boldsymbol{G}_{add}^{-1} \bar{\boldsymbol{\mu}}}{\bar{\boldsymbol{\mu}}' \boldsymbol{G}_{add}^{-1} \bar{\boldsymbol{\mu}}}$ et $p_f = 1 - \boldsymbol{p}'\mathbb{1}$ minimisent le GMD d'un portefeuille composé d'actifs risqués et d'un actif sans risque afin d'atteindre un excès de rendement cible $\widetilde{\mu}_p^0$. Il est le résultat du problème de minimisation suivant :*

$$\begin{aligned} \min \Gamma^P &= \min \boldsymbol{p}'\boldsymbol{G}\boldsymbol{p} \\ s.c.\ \mu_p^p &= \widetilde{\mu}_p^0. \end{aligned}$$

Démonstration. Notons que $\mathbf{p}'\mathbb{1} = 1$ n'est pas une contrainte parce-que la richesse dont nous disposons n'est pas obligatoirement et entièrement allouée aux actifs risqués ; une partie de la richesse peut être allouée dans l'actif sans risque. Le lagrangien s'écrit :

$$L(\mathbf{p}, \lambda) = \mathbf{p}'\mathbf{G}\mathbf{p} + \lambda(\mathbf{x}'\bar{\boldsymbol{\mu}} - \widetilde{\mu}_p^0)$$

Les conditions de premier ordre pour un minimum sont :

$$\frac{\partial L(\mathrm{p}, \lambda)}{\partial \mathbf{p}} = \mathbf{G}_{add}\,\mathbf{p} + \lambda\bar{\boldsymbol{\mu}} = 0 \tag{4.15}$$

$$\frac{\partial L(\mathrm{p}, \lambda)}{\partial \mathbf{p}} = \mathbf{x}'\bar{\boldsymbol{\mu}} - \widetilde{\mu}_p^0 = 0 \tag{4.16}$$

En utilisant (4.15), il est possible de déduire $\mathbf{p}$ en fonction de λ :

$$\begin{aligned} \mathbf{G}_{add}\,\mathbf{p} &= -\lambda\,\bar{\boldsymbol{\mu}} \\ \mathbf{p} &= -\lambda\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}} \end{aligned} \tag{4.17}$$

L'équation (4.16) implique que $\mathbf{x}'\bar{\boldsymbol{\mu}} = \mathbf{x}\bar{\boldsymbol{\mu}}' = \widetilde{\mu}_p^0$. En multipliant (4.17) par $\bar{\boldsymbol{\mu}}'$:

$$\bar{\boldsymbol{\mu}}'\,\mathbf{p} = -\lambda\,\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}} = \widetilde{\mu}_p^0,$$

de laquelle, l'on peut tirer λ :

$$\lambda = -\frac{\widetilde{\mu}_p^0}{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}} \tag{4.18}$$

En remplaçant (4.18) dans (4.17), on obtient la solution de $\mathbf{p}$:

$$\begin{aligned} \mathbf{p} &= -\lambda\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}} \\ &= -(-\frac{\widetilde{\mu}_p^0}{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}})\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}} \\ \mathbf{p} &= \widetilde{\mu}_p^0\,\frac{\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}} \end{aligned} \tag{4.19}$$

La solution pour p_f est donc $p_f = 1 - \mathbf{p}'\mathbb{1}$.

□

Concernant le portefeuille tangent [6] $\mathbf{t}$, la totalité de la richesse est investie en actifs risqués, alors $\mathbf{t}'\mathbb{1} = \mathbb{1}'\mathbf{t} = 1$.

Theorem 4.2.4. *Le vecteur* $\boldsymbol{t} = \frac{\boldsymbol{G}_{add}^{-1}(\mu - r_f\mathbb{1})}{\mathbb{1}'\boldsymbol{G}_{add}^{-1}(\mu - r_f\mathbb{1})}$ *est le vecteur des poids du portefeuille tangent.*

Démonstration. En remplaçant $\mathbf{p}$ par $\mathbf{t}$ dans l'équation (4.19), nous avons :

$$\begin{aligned} \mathbf{t} &= \widetilde{\mu}_p^t \, \frac{\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}} \\ \mathbb{1}'\mathbf{t} &= \widetilde{\mu}_p^t \, \frac{\mathbb{1}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}} \\ 1 &= \widetilde{\mu}_p^t \, \frac{\mathbb{1}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}, \end{aligned} \tag{4.20}$$

laquelle implique que :

$$\widetilde{\mu}_p^t = \frac{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\mathbb{1}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}} \tag{4.21}$$

En remplaçant (4.21) dans l'expression de t, nous avons une solution explicite pour le vecteur poids t du portefeuille tangent :

$$\begin{aligned} \mathbf{t} &= \left(\frac{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\mathbb{1}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}\right) \frac{\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\bar{\boldsymbol{\mu}}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}} \\ &= \frac{\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}}{\mathbb{1}'\,\mathbf{G}_{add}^{-1}\,\bar{\boldsymbol{\mu}}} \\ \mathbf{t} &= \frac{\mathbf{G}_{add}^{-1}\,(\boldsymbol{\mu} - r_f\mathbb{1})}{\mathbb{1}'\,\mathbf{G}_{add}^{-1}\,(\boldsymbol{\mu} - r_f\mathbb{1})} \end{aligned} \tag{4.22}$$

□

6. Le portefeuille tangent est un portefeuille efficient qui tient compte en plus des actifs risqués, de la présence d'actif sans risque dans l'univers des titres disponibles. Le portefeuille tangent est le portefeuille efficient le mieux diversifié. En présence d'un actif sans risque, tout investisseur rationnel devrait détenir le portefeuille tangent.

Remarques :

— Si le rendement de l'actif sans risque, r_f, est plus petit que l'espérance de rendement du portefeuille à risque (GMD) minimal global (PRMG), le portefeuille tangent (ou de marché) a une pente positive.

— Si le rendement de l'actif sans risque, r_f, est égal à l'espérance de rendement du portefeuille à risque (GMD) minimal global (PRMG), le portefeuille tangent (ou de marché) n'est pas défini.

— Si le rendement de l'actif sans risque, r_f, est plus grand que l'espérance de rendement du portefeuille à risque (GMD) minimal global (PRMG), le portefeuille tangent (ou de marché) a une pente négative.

Simulations

Cette partie permet de généraliser les résultats obtenus ci-dessus. Trois sous-parties peuvent être observées. Premièrement, il est simulé la dynamique du prix d'un actif et d'un portefeuille d'actifs grâce à des mouvements browniens géométriques. Deuxièmement, une simulation du portefeuille à risque minimal global (PRMG) et de la frontière efficiente est faite à partir des résultats obtenus dans la première partie. Troisièmement, une simulation de Monte Carlo est faite afin de mettre en évidence la robustesse du moyenne-GMD avec des données contaminées à travers le calcul des erreurs quadratiques moyennes (MSE en anglais) des poids du portefeuille à risque minimal global.

Le processus de modélisation du prix d'un actif

Des travaux de Black et Scholes, la modélisation retenue pour représenter la dynamique des prix d'un actif, avec S le cours de l'actif, est l'utilisation des mouvements browniens géométriques de type :

$$dS = rSdt + \widehat{\sigma}Sdz, \tag{4.23}$$

où r_f et $\widehat{\sigma}$ représentent respectivement le taux sans risque et la volatilité selon le GMD de l'actif sous-jacent, et où z suit un processus de Wiener. Ce type de processus continu doit être discrétisé préalablement en N intervalles de longueur $\Delta t = T/N$ où T représente la période de temps considérée avec $t \in \{1, \ldots, T\}$. Pour réécrire l'équation (4.23), il est nécessaire d'utiliser une approximation d'Euler d'une diffusion :

$$S(t+\Delta t) - S(t) = r_f S(t)\Delta t + \widehat{\sigma} S(t)\epsilon + \sqrt{\Delta t}, \tag{4.24}$$

où $\epsilon \sim \mathcal{N}(0,1)$, est un nombre aléatoire qui suit une loi normale standard. Pour la simulation, il est utilisé $\ln(S)$ plutôt que S. En applicant le lemme d'Itô, la dynamique de $\ln(S)$ s'écrit :

$$d\ln(S) = (r_f - \frac{\widehat{\sigma}^2}{2})dt + \widehat{\sigma} dz, \tag{4.25}$$

et qui peut être discrétisé à l'aide de cette approximation :

$$\ln(S(t+\Delta t)) - \ln(S(t)) = (r_f - \frac{\widehat{\sigma}^2}{2})\Delta t + \widehat{\sigma}\epsilon\sqrt{(\Delta t)} \tag{4.26}$$

En passant chacun des membres de l'équation (4.26) à l'exponentielle, l'on obtient :

$$S(t+\Delta t) = S(t)\ \exp\{\ (r_f - \frac{\widehat{\sigma}^2}{2})\Delta t + \widehat{\sigma}\epsilon\sqrt{(\Delta t)}\ \} \tag{4.27}$$

Cette modélisation est adaptée au processus des prix, également au processus des indices boursiers mais innapropriée à la modélisation de la dynamique des taux d'intérêts.

Le processus de modélisation d'un portefeuille d'actifs : un exemple

Il est nécessaire de pouvoir générer des trajectoires de prix d'actifs corrélés afin de simuler la valeur d'un portefeuille et mesurer l'impact de la diversification. Dans cette partie, pour l'illustration, il est fait la simulation de quatre actifs (mouvements browniens géométriques) corrélés. Il est possible de simuler les trajectoires de prix de quatre actifs à partir de quatre nombres aléatoires indépendants $\eta_1, \eta_2, \eta_3, \eta_4$. Le principe est de bruiter judicieusement le premier nombre généré à l'aide des trois suivants afin d'obtenir les niveaux de corrélation souhaités. Soient $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ les quatre termes aléatoires à utiliser pour discrétiser les mouvements browniens et les corréler :

$$\begin{cases} \epsilon_1 = \eta_1 \\ \epsilon_2 = a\,\eta_1 + b\,\eta_2 \\ \epsilon_3 = c\,\eta_1 + d\,\eta_2 + e\,\eta_3 \\ \epsilon_4 = f\,\eta_1 + g\,\eta_2 + h\,\eta_3 + i\,\eta_4 \end{cases}$$

Ce système d'équations peut être écrit sous la forme matricielle suivante :

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a & b & 0 & 0 \\ c & d & e & 0 \\ f & g & h & i \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{pmatrix}$$

Pour cela, il est nécessaire d'utiliser la décomposition de Cholesky de la matrice des corrélations des taux de rentabilité des actifs qui composent le portefeuille afin d'obtenir les inconnues (a, b, c, d, e, f, g, h, i).
Soit $\mathbf{\Psi}$ la matrice de corrélation des rentabilités des 4 actifs telle que :

$$\mathbf{\Psi} = \begin{pmatrix} 1 & -0.5 & 0.25 & 0.5 \\ -0.5 & 1 & -0.75 & 0 \\ 0.25 & -0.75 & 1 & -0.25 \\ 0.5 & 0 & -0.25 & 1 \end{pmatrix}$$

Par application de la décomposition de Cholesky de la matrice $\mathbf{\Psi}$, l'on obtient une matrice $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -0.5 & 0.866 & 0 & 0 \\ 0.25 & -0.721 & 0.645 & 0 \\ 0.5 & 0.288 & -0.258 & 0.774 \end{pmatrix}$$

Afin d'obtenir les 4 nombres aléatoires $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ selon les corrélations souhaitées, il suffit de simuler 4 nombres aléatoires indépendamment distribués qui suivent une loi normale pour chacun des 250 jours (année de trading) que nous multiplions ensuite par la matrice $\mathbf{A}$. Le taux de rentabilité r_i pour l'actif i est alors :

$$r_i = (\mu_i - \frac{\widehat{\sigma}_i^2}{2})\Delta t + \widehat{\sigma}_i \epsilon_i \sqrt{\Delta t}, \tag{4.28}$$

avec μ_i et $\widehat{\sigma}_i$ qui représentent respectivement l'espérance de rentabilité et la volatilité annuelle selon le GMD de l'actif i et Δt qui est égal à $1/250$. Pour éviter une simulation fastidieuse, l'on suppose un taux de rentabilité de 0.8 et une volatilité de 0.025 pour l'ensemble des 4 actifs avec un prix initial fixé à 100.

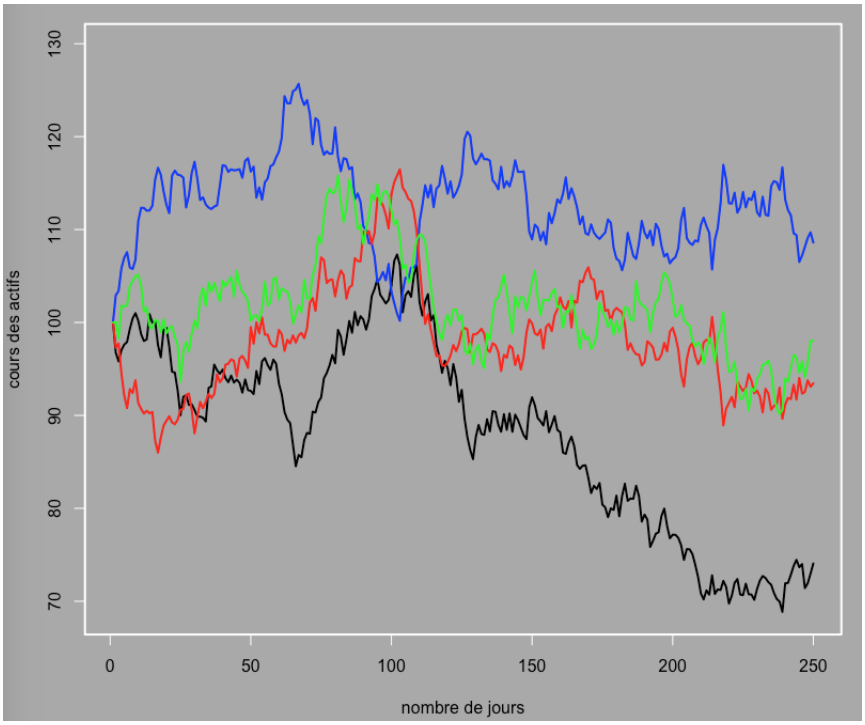


FIGURE 4.1 – Simulation du cours de quatre actifs corrélés

La frontière efficiente selon le critère moyenne-GMD

La théorie de la frontière efficiente consiste à sélectionner des portefeuilles optimaux en prenant en compte à la fois la rentabilité espérée et l'incertitude sur cette rentabilité. L'incertitude (ou risque) est mesurée par le GMD. Après avoir simuler les actifs constituants notre portefeuille et en s'appuyant sur les Théorèmes (4.2.1) et (4.2.2), il est simulé et illustré une frontière efficiente selon le critère moyenne-GMD.

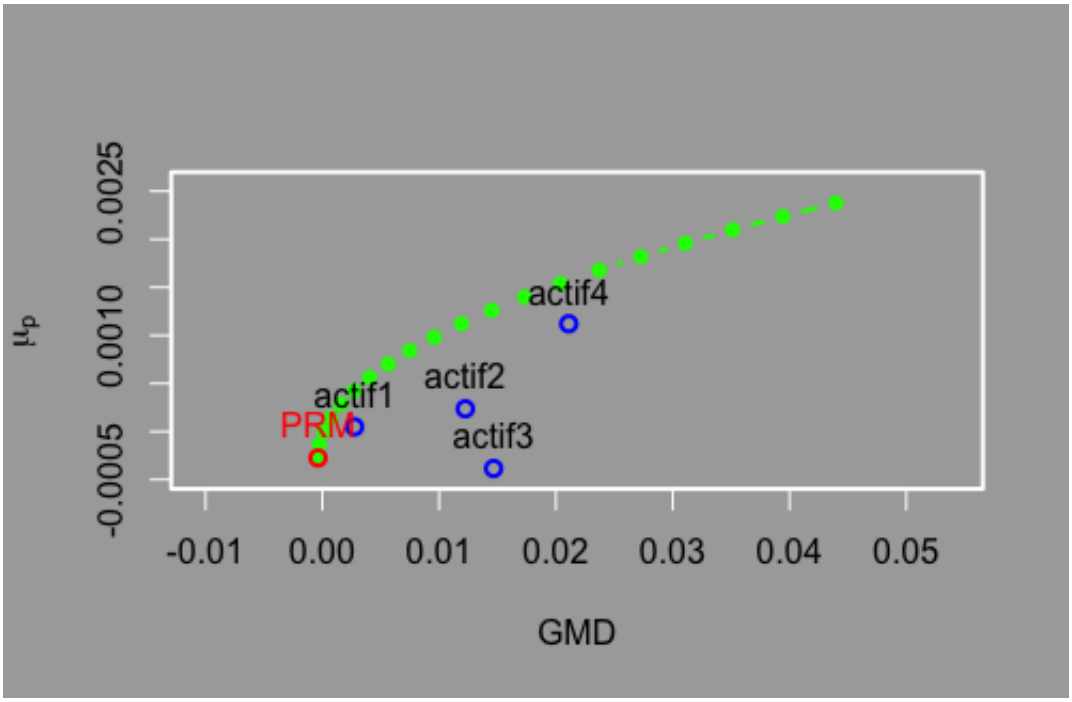


FIGURE 4.2 – Simulation d'une frontière efficiente avec le GMD

Simulation de Monte Carlo

Une expérience (ou simulation) de Monte Carlo est menée avec 4 mouvements browniens géométriques corrélés (N = 250 jours [7]) qui décrivent la dynamique des prix d'actifs permettant de constituer un portefeuille d'actifs. La matrice que forme ces 4 mouvements browniens géométriques corrélés est notée **X**. Cette expérience est faite afin d'évaluer la qualité des poids p_{min} du portefeuille à risque minimal global (PRMG) au moyen de l'estimation des MSE. Une contamination aléatoire (1000 différentes contaminations) de la base de données de départ des actifs du portefeuille est faite en augmentant l'intensité de la contamination de 1 à 1000. Les MSE des poids p_{min} du PRMG se calculent de la manière suivante :

$$MSE_{p_{min}} = \frac{\sum_{i=1}^{1000}(p_{min}^{oi} - p_{min}^{k})^2}{1,000}, \ \forall k = 1, \ldots, 4, \tag{4.29}$$

7. une année de trading

avec p_{min}^{oi} qui le poids minimal d'un actif avec les données contaminées et p_{min}^{k} le poids minimal d'un actif k sans contamination des données.

Algorithm 5: Simulation de Monte Carlo du PRMG

Result: Pondération robuste en Gini avec contamination des données

$\theta = 1$ [θ est la valeur de l'outlier] et $N = 250$;
repeat
    Générer 4 mouvements browniens géométriques corrélés $\mathbf{X}$;
    Contaminer 1 observation (ligne) de $\mathbf{X}$ en la multipliant par θ [choix aléatoire de la ligne] ;
    Calculer les poids p_{min} du PRMG ;
until $\theta = 1000$ [par pas de 1];
return Mean squared Errors (MSE) des poids p_{min} du PRMG ;

En absence d'outliers, les pondérations de chaque actif constituant le PRMG sont quasiment identiques quelque soit la méthode (Moyenne - GMD ou Moyenne - Variance). Le tableau II montre que les MSE issus de la pondération du PRMG selon le modèle Moyenne - GMD sont nettement moins importants que ceux issus du modèle Moyenne - Variance. Ce tableau met donc en évidence la robustesse du modèle Moyenne - GMD pour la construction du PRMG en présence d'outliers dans les données.

	Poids avant contamination		Poids moyens après contamination	
	Gini	Variance	Gini	Variance
actif 1	**0.15266812**	0.15533875	**0.15281184**	0.3504279
actif 2	**0.41872857**	0.41616340	**0.41890532**	0.9184239
actif 3	**0.36072678**	0.36235120	**0.36080477**	0.6623500
actif 4	**0.06787654**	0.06614665	**0.06747808**	-0.9312018

TABLE 4.1 – Les poids avant contamination et les poids moyens après contamination du portefeuille à risque minimal global

	MSE Gini	MSE Variance
actif 1	**4.755394e-05**	**772.6070**
actif 2	**1.767764e-05**	**237.8956**
actif 3	**2.213425e-05**	**344.8228**
actif 4	**4.655301e-05**	**662.8443**

TABLE 4.2 – Les MSE des poids du portefeuille à risque minimal global

Le modèle d'évaluation des actifs financiers

Dans cette section il est présenté les résultats théoriques des modèles d'évaluation d'actifs financiers avant de faire une approche par des simulations.

Résultats théoriques

Le Modèle d'Evaluation Des Actifs Financiers propose une explication rigoureuse de la formation des prix à l'équilibre sur les marchés des capitaux et constitue un apport important à la théorie financière. Ce modèle fut développé par Sharpe et amélioré par Lintner et Mossin dans les années 60 et montre qu'à l'équilibre, seul le risque non diversifiable est rémunéré. Les hypothèses principales du modèle sont les suivantes :

Hypothèses 4.2.1. – **(H1)** – : *Il existe une relation croissante entre rendement et risque : les investisseurs exigent une rentabilité d'autant plus importante que le risque est élevé.*

Hypothèses 4.2.2. – **(H2)** – : *Il existe un actif sans risque disponible sur le marché.*

Hypothèses 4.2.3. – **(H3)** – : *Tous les investisseurs disposent des mêmes informations et donc ont les mêmes anticipations de rentabilité et les mêmes mesures de risque.*

De ces hypothèses, l'espérance de la rentabilité d'un titre risqué pris isolément R_i (ou d'un portefeuille R_p) est fonction de l'actif sans risque r_f et de l'espérance de la rentabilité du portefeuille de marché R_M (portefeuille détenu par tous les investisseurs) :

$$\mathbb{E}(R_p) = r_f \ + \ \beta_i \left[\mathbb{E}(R_M) - r_f\right] \text{ avec } \beta_i = \frac{Cov(R_p, R_M)}{\sigma^2(R_M)}, \tag{4.30}$$

avec $\sigma()$ la volatilité de l'actif risqué ou du portefeuille. Comme nous l'avons montré au Chapitre 1, l'estimateur β repose sur l'hypothèse suivante :

Hypothèses 4.2.4. – **(H4)** – : *Les chroniques R_{pt} et R_{Mt} ne présentent aucune donnée aberrante (outlier).*

En présence de valeurs extrêmes (krach boursier par exemple), l'hypothèse **H4** est violée. Ce qui justifie l'utilisation de l'indice de Gini. La droite de marché peut alors être estimée par régression Gini lorsque des outliers sont présents dans les données :

$$\mathbb{E}(R_p) = r_f \ + \ \beta_p^G \left[\mathbb{E}(R_M) - r_f\right] \text{ avec } \beta_p^G = \frac{Cog(R_p, R_M)}{Cog(R_M, R_M)}. \tag{4.31}$$

La droite de marché de l'action i estimée par régression Gini est :

$$\mathbb{E}(R_i) = r_f \ + \ \beta_i^G \left[\mathbb{E}(R_M) - r_f\right] \text{ avec } \beta_i^G = \frac{Cog(R_i, R_M)}{Cog(R_M, R_M)}. \tag{4.32}$$

La prime de risque ρ_i, apparait comme suit :

$$\rho_i = \mathbb{E}(R_i) - r_f = \beta_i^G \left[\mathbb{E}(R_M) - r_f\right]. \tag{4.33}$$

Plus les investisseurs sont exposés au portefeuille de marché, plus leur rendement moyen à long terme sera théoriquement important car ils auront pris plus de risque. Le MEDAF montre que le bêta (β^G) mesure l'exposition ou la sensibilité de l'actif risqué ou du portefeuille au risque de marché (ou au portefeuille de marché). Etant donné que $\beta_i^G = Cog(R_i, R_M)/Cog(R_M, R_M)$, alors le bêta du marché, $\beta_M = Cog(R_M, R_M)/Cog(R_M, R_M) = 1$. Deux cas peuvent être distingués :
– Cas 1. $\beta_i^G > 1$: le titre (ou portefeuille) i est dit "agressif" et indique que ce titre est plus risqué que le marché.
– Cas 2. $\beta_i^G < 1$: le titre (ou portefeuille) i est dit "défensif" et indique que ce titre est moins risqué que le marché.

Deux importantes conclusions se dégagent du MEDAF. La première est que tout investisseur rationnel devrait détenir à la fois l'actif sans risque et le portefeuille de marché. En fonction du degré d'aversion au risque, les proportions d'actif sans risque et de portefeuille de marché sont modulées. Un investisseur plus averse au risque qu'un autre aura une proportion de sa richesse investie dans l'actif sans risque plus importante qu'un autre. La seconde qui est très fondamentale montre que seul le risque non diversifiable ou risque systématique (bêta) peut être rémunéré. Le risque spécifique à chaque entreprise peut facilement être éliminé par un gestionnaire habile en construisant un portefeuille bien diversifié.

Le MEDAF a subi un certain nombre de critiques quant aux hypothèses restrictives mais nécessaires pour sa construction. De plus, pour la mise en oeuvre du modèle, il est difficile de déterminer correctement le portefeuille de marché qui est souvent réduit à l'indice de marché. Selon Roll (1977), le MEDAF n'est pas testable car la rentabilité du marché ne peut être évaluée qu'en référence à un indice de marché [8]. En outre, la critique la plus importante est celle de Fama et French (1992) qui concluent leurs travaux par le fait que seul le rendement du marché ne peut correctement expliquer le rendement moyen à long terme d'un actif risqué ou d'un portefeuille. En vue de faire face à cette critique majeure, plusieurs modèles multifactoriels voient le jour dont le modèle d'évaluation par arbitrage (MEA).

Le modèle d'évaluation par arbitrage

Résultats théoriques

Le modèle d'évaluation par arbitrage (MEA ou APT en anglais) est une théorie qui généralise celle du MEDAF. Ce modèle suppose que les cours des actifs risqués sont influencés par un nombre limité de facteurs communs à l'ensemble des actifs (comme l'inflation, les prix du pétrole, etc.) et par un facteur spécifique à chaque actif. Ce dernier est totalement indépendant des autres facteurs de risque. Cette hypothèse s'écrit :

$$r_{it} = \alpha_i + \beta_{1i}.f_{1t} + \beta_{2i}.f_{2t} + \cdots + \beta_{ki}.f_{kt} + \epsilon_{it}, \tag{4.34}$$

8. L'indice phare de la place financière ou l'indice de marché n'intègre pas tous les actifs risqués du marché considéré, alors que théoriquement il est défini par tous les titres existants.

où r_{it} est le taux de rendement (ou rentabilité) de l'actif i sur la période t, f_{1t} à f_{kt} les k facteurs communs dont l'un pouvant être celui du marché, β_{1i} à β_{ki} les sensibilités ou expositions de l'actif i à chaque facteur de risque commun. La constante α_i est spécifique à l'actif i et ϵ_{it} [9] est le facteur spécifique à l'actif i.

La version du MEA robuste aux outliers est, grâce à la régression Gini,

$$r_{it} = \alpha_i^G + \beta_{1i}^G.f_{1t} + \beta_{2i}^G.f_{2t} + \cdots + \beta_{ki}^G.f_{kt} + \epsilon_{it}^G \tag{4.35}$$

Le MEA permet aux investisseurs et aux théoriciens de privilégier un petit nombre de facteurs communs à tous les titres. Toutes les autres sources de risques sont spécifiques et peuvent donc être éliminées par diversification. A partir d'un raisonnement par arbitrage, cette première hypothèse nous permet de montrer que dans un marché efficient [10], le rendement moyen à long terme (ou rentabilité anticipé) pour chaque titre est une combinaison linéaire des bêtas relatifs à chaque facteur. Il s'écrit comme suit :

$$\mathbb{E}(r_i) = \alpha_i^G + \lambda_1.\beta_{1i}^G + \lambda_2.\beta_{2i}^G + \cdots + \lambda_k.\beta_{ki}^G. \tag{4.36}$$

Rappelons qu'à l'équilibre, seuls les risques systématiques sont rémunérés et non les risques spécifiques. Il y a par conséquent, k facteurs communs de risque (non diversifiables ou systématiques) et k primes de risque (λ_1 à λ_k) liées à chaque facteur de risque, avec λ_0 le taux d'intérêt sans risque r_f lorsqu'il existe. L'équation précédente pourrait se réécrire donc :

$$\mathbb{E}(r_i) = r_f + \lambda_1.\beta_{1i}^G + \lambda_2.\beta_{2i}^G + \cdots + \lambda_k.\beta_{ki}^G. \tag{4.37}$$

Le modèle MEA reste muet sur la nature des différents facteurs. Ceci est à la fois un bien et un mal. Un bien pour les gestionnaires d'actifs et les traders car ils auront la liberté de choisir eux-mêmes les facteurs qui leur semblent important mais un mal pour le théoricien à qui le choix des facteurs pose un réel problème. L'une des méthodes MEA est d'utiliser les facteurs macroéconomiques (inflation, chômage, taux d'intérêts directeurs) susceptibles d'influencer le cours des actifs. A partir de cette méthode, les chercheurs ont fait le constat suivant : l'indice phare d'une place financière ou boursière (exemple l'indice de Cotation Assistée en Continu des 40 premières

9. Le terme spécifique ϵ_{it} a les caractéristiques d'un bruit blanc. Il est différent pour chaque observation t mais nul en moyenne.

10. Par marché efficient, il est sous-entendu dans ce cas, que toutes les opportunités de pur arbitrage sont très vite éliminées.

valeurs mobilières de Paris CAC40) n'a pas d'influence sur le rendement des actifs de cette place boursière. En effet, le rendement moyen à long terme de l'indice phare d'une place financière est entièrement déterminé par les risques macroéconomiques [11].

Une autre des méthodes les plus couramment utilisées consiste à utiliser des méthodes statistiques, comme l'Analyse en Composantes Principales sur l'univers des rentabilités pour en extraire les facteurs communs non corrélés entre eux. Les facteurs dérivés sont des rendements de portefeuilles d'actifs et engloblent tous les facteurs macroéconomiques, alors qu'en utilisant des facteurs macroéconomiques spécifiques, on peut en omettre certains. Le principal inconvénient est que les facteurs latents issus de l'ACP ne sont pas susceptibles d'interprétation économique. On peut néanmoins recourir aux contributions absolues (CTA) et contributions relatives (CTR) définies dans les Chapitres précédents.

Le modèle MEA peut être utilisé dans de nombreux cas où le MEDAF est utilisé comme pour mesurer la rentabilité ajustée du risque dans l'évaluation d'un projet d'investissement. Ce modèle est souvent utilisé également pour la mesure de la performance des portefeuilles (section suivante). Le modèle MEA permet en effet dans ce cas au praticien de déterminer les facteurs qui ont eu une influence sur la performance de son portefeuille.

Simulations

Une expérience (ou simulation) de Monte Carlo est menée sur les primes de risque issues du Modèle d'Evaluation par Arbitrage classique (MEA) et du Modèle d'Evaluation par Arbitrage avec une approche Gini (MEA-Gini). Nous utilisons 4 mouvements browniens géométriques qui décrivent la dynamique des prix d'actifs permettant de constituer un portefeuille d'actifs et 4 facteurs corrélés aux actifs. Cette expérience est faite afin d'évaluer la qualité et la robustesse des primes de risque du MEA-Gini au moyen de l'estimation des MSE. Une contamination aléatoire de 50 lignes (100 différentes contaminations) de la base de données de départ des facteurs est faite avec une variable aléatoire suivant une loi normale. Les MSE des primes de risque $P.R$ se calculent de la manière suivante :

11. Source : Massoud Mussavian "Evaluation des risques : l'alternative du modèle APT.", L'Art de la Finance, Les Echos – Vendredi 27 et Samedi 28 Mars 1998.

$$MSE_{P.R} = \frac{\sum_{i=1}^{100}(P.R^{oi} - P.R^{k})^2}{1,00}, \ \forall k = 1, \ldots, 2, \tag{4.38}$$

avec $P.R^{oi}$ qui est la prime de risque calculée avec les données contaminées et $P.R^{k}$ la k-ième prime de risque sans contamination des données.

Algorithm 6: Simulation de Monte Carlo des Primes de risque du MEA

Result: Prime de risque robuste en Gini avec contamination des données

$\theta = 1$ [θ est la valeur de l'outlier] et $N = 250$;
repeat
  Générer 4 mouvements browniens géométriques corrélés $\mathbf{X}$;
  Générer 4 facteurs corrélés à $\mathbf{X}$;
  Générer un vecteur aléatoire de 50 entiers compris entre 1 et 250 ;
  Contaminer 50 observations (lignes) des facteurs en la multipliant par un nombre aléatoire [choix aléatoire des lignes] ;
  Calculer les primes de risque $P.R$ issues du modèle d'évaluation par arbitrage ;
until $\theta = 100$ [par pas de 1];
return Mean squared Errors (MSE) des primes de risque $P.R$ du MEA ;

Les primes de risque des différentes approches des modèles d'évaluation par arbitrage (MEA & MEA-Gini) sont nettement différentes avant la contamination avec de faibles primes de risques dans l'ensemble. Les primes de risque moyennes après contamination sont assez proches à l'exception de la seconde prime par l'approche classique. Les primes de risque moyennes après contamination des deux approches sont identiques pour la prime de risque 1.

	PR avant contamination		PR moyennes après contamination	
	Gini	Variance	Gini	Variance
Prime 1	**1.12e-03**	9.69e-04	**-0.302**	-0.302
Prime 2	**2.00e-05**	1.54e-05	**-0.304**	0.286

TABLE 4.3 – Les primes de risque avant contamination et les primes de risque moyennes après contamination du MEA

Bien que les MSE de la première prime de risque soient identiques, il existe une énorme différence au niveau des MSE de la seconde prime de risque. Le tableau ci-dessous met donc en évidence la robustesse des primes de risque du modèle d'évaluation par arbitrage par une approche Gini en présence d'outliers dans les données.

	MSE Gini	MSE Variance
Prime de risque 1	**1.28**	**1.28**
Prime de risque 2	**1.28**	**16.02**

TABLE 4.4 – Les MSE des primes de risque du modèle d'évaluation par arbitrage

Le ratio (ou mesure) de performance généralisé de Treynor

Résultats théoriques

La mesure de performance d'un portefeuille ou d'un investissement a pour objectif d'évaluer le résultat obtenu par rapport à d'autres stratégies d'investissement comparables. Dans cette partie, nous allons revisiter le ratio de performance de Treynor avant d'introduire la généralisation de celui-ci grâce au modèle MEA-Gini.

Le ratio de performance de Treynor fut créé par Jack Treynor (1965) et est basé sur le MEDAF. Ce ratio représente le rapport entre l'excès du rendement du portefeuille (ou de l'investissement) vis-à-vis du marché et son risque systématique, déterminé par le bêta issu du MEDAF :

$$RT_p = \frac{\mathbb{E}(r_p) - r_f}{\beta_p}, \tag{4.39}$$

avec RT_p le ratio de performance de Treynor du portefeuille. La version robuste aux outliers de cet indicateur est :

$$RT_p^G = \frac{\mathbb{E}(r_p) - r_f}{\beta_p^G}. \tag{4.40}$$

Le ratio de performance de Treynor d'un portefeuille permet d'évaluer la rentabilité d'un portefeuille par rapport au risque engagé. Ainsi plus RT_p^G est élevé, plus ce portefeuille offre une rentabilité intéressante par rapport au risque encouru (ou engagé).

Dans le cas du ratio de performance généralisé de Treynor (RTG_p), il s'agit d'une configuration multi-indice du ratio de Treynor original qui conserve les mêmes propriétés géométriques et analytiques que celui-ci. Il doit s'agir essentiellement, d'une fonction monotone du risque systématique du portefeuille, être comparable à un portefeuille de référence (Benchmark), fournir un classement semblable pour les portefeuilles avec le même risque, être indépendant de l'échelle des primes de risque et fournir des mesures de performance indépendantes du choix du modèle. Dans le cas d'un multi-index linéaire, ces exigences sont remplies en normalisant la prime de risque, en utilisant un benchmark et en obtenant un hyperplan orthonormé grâce aux facteurs.

Le RTG_p^G est défini comme l'excès de rendement du portefeuille par unité de moyenne pondérée de la prime de risque systématique, normalisé par la moyenne pondérée de la prime de risque systématique d'un benchmark. Par des simulations numériques, Hübner (2005) montre que les classements de portefeuilles produits avec cette mesure sont plus précis et plus stables que ceux fournis par l'alpha de Jensen et le ratio d'information. Cette mesure de performance est basée sur le MEA. Rappelons que selon le modèle MEA, la rentabilité anticipée d'un portefeuille $\mathbb{E}(r_p)$ s'écrit :

$$\begin{aligned}
\mathbb{E}(r_p) &= r_f + \tilde{\mu}_1.\beta_{1p}^G + \tilde{\mu}_2.\beta_{2p}^G + \cdots + \tilde{\mu}_k.\beta_{kp}^G \\
\mathbb{E}(r_p) - r_f &= \tilde{\mu}_1.\beta_{1p}^G + \tilde{\mu}_2.\beta_{2p}^G + \cdots + \tilde{\mu}_k.\beta_{kp}^G \\
\mathbb{E}(r_p) - r_f &= \sum_{j=1}^{k} \beta_{pj}^G.\tilde{\mu}_j
\end{aligned}$$

Le ratio de performance généralisé de Treynor, robuste aux outliers, s'écrit comme suit :

$$\begin{aligned} RTG_p^G &= \frac{\mathbb{E}(r_p) - r_f}{\frac{\sum_{j=1}^{k} \beta_{pj}^{G*} w_j}{\sum_{j=1}^{k} w_j}} \\ RTG_p^G &= (\mathbb{E}(r_p) - r_f) \, . \, \frac{\sum_{j=1}^{k} w_j}{\sum_{j=1}^{k} \beta_{pj}^{G*} w_j} \end{aligned}$$

où $j = 1, \cdots, k$ désigne le nombre de primes de risque distinctes, $w_j = \tilde{\mu}_j \beta_{mj}^G$ avec $\beta_{pj}^{G*} = \beta_{m1}^G, \cdots, \beta_{mk}^G$ les sensibilités du portefeuille de référence aux différents risques systématques et $\beta_{pj}^{G*} = \frac{\beta_{pj}^G}{\beta_{mj}^G}$ avec $\beta_{pj}^G = \beta_{p1}^G, \cdots, \beta_{pk}^G$ les sensibilités du portefeuille aux différents risques systématiques. Alors l'équation est réécrite comme suit :

$$RTG_p^G = (\mathbb{E}(r_p) - r_f) \, . \, \frac{\sum_{j=1}^{k} \tilde{\mu}_j \beta_{mj}}{\sum_{j=1}^{k} \frac{\beta_{pj}^G}{\beta_{mj}^G} \tilde{\mu}_j \beta_{mj}} \tag{4.41}$$

Simulations

Une expérience (ou simulation) de Monte Carlo est menée sur les Ratios de Treynor Généralisés RTG à partir des sensibilités issues du Modèle d'Evaluation par Arbitrage classique (MEA) et du Modèle d'Evaluation par Arbitrage avec une approche Gini (MEA-Gini). Nous utilisons 4 mouvements browniens géométriques qui décrivent la dynamique des prix d'actifs permettant de constituer un portefeuille d'actifs et 4 facteurs corrélés aux actifs. Cette expérience est faite afin d'évaluer la qualité et la robustesse des Ratios de Treynor Généralisés avec une approche Gini au moyen de l'estimation des MSE. Une contamination aléatoire de 50 lignes (100 différentes contaminations) de la base de données de départ des facteurs est faite avec une variable aléatoire suivant une loi normale. Les MSE des ratios de Treynor généralisés RTG se calculent de la manière suivante :

$$MSE_{RTG} = \frac{\sum_{i=1}^{100} (RTG^{oi} - RTG^k)^2}{1,00}, \ \forall k = 1, \ldots, 4, \tag{4.42}$$

avec RTG^{oi} qui est le ratio de Treynor généralisé calculé avec les données contaminées et RTG^k le k-ième ratio de Treynor généralisé sans contamination des données.

Algorithm 7: Simulation de Monte Carlo des Ratios de Treynor Généralisés

Result: Prime de risque robuste en Gini avec contamination des données

$\theta = 1$ [θ est la valeur de l'outlier] et $N = 250$;

repeat

  Générer 4 mouvements browniens géométriques corrélés $\mathbf{X}$;

  Générer 4 facteurs corrélés à $\mathbf{X}$;

  Générer un vecteur aléatoire de 50 entiers compris entre 1 et 250 ;

  Contaminer 50 observations (lignes) des facteurs en la multipliant par un nombre aléatoire [choix aléatoire des lignes] ;

  Calculer les Ratios de Treynor Généralisés RTG à partir des sensibilités du modèle d'évaluation par arbitrage ;

until $\theta = 100$ [par pas de 1];

return Mean squared Errors (MSE) des Ratios de Treynor Généralisés RTG ;

Les Ratios de Treynor Généralisés des différentes approches selon les modèles d'évaluation par arbitrage (MEA $\&$ MEA-Gini) sont nettement différents. Avant la contamination, les ratios issus de l'approche Gini sont faibles mais tous positifs tandis que ceux issus de l'approche Variance sont pour la plupart négatif. Après la contamination, les deux approches donnent des RTG moyens différents avec notamment ceux de l'approche Variance nuls et des RTG moyens négatifs pour les actifs 1 $\&$ 3 par l'approche Gini.

	RTG avant contamination		RTG moyens après contamination	
	Gini	Variance	Gini	Variance
Actif 1	**2.14e-04**	-0.57	**-9.07**	0.00
Actif 2	**1.24e-04**	0.04	**0.08**	0.00
Actif 3	**1.07e-04**	-0.06	**-0.02**	0.00
Actif 4	**1.07e-04**	-0.05	**0.1**	0.00

TABLE 4.5 – Les RTG avant contamination et les RTG moyens après contamination

Le tableau ci-dessous montre que les MSE du Ratio de Treynor Généralisé des actifs (simulés) issus l'approche Gini sont moins importants que ceux issus de l'approche classique. Le tableau 4.6 met donc en évidence la robustesse du Ratio de Treynor Généralisé avec l'approche Gini en présence d'outliers dans les données.

	MSE Gini	MSE Variance
RTG.actif 1	**1.33e-07**	**1.17e+04**
RTG.actif 2	**2.8e-07**	**6.08e-01**
RTG.actif 3	**1.42e-07**	**5.62e-01**
RTG.actif 4	**1.12e-06**	**4.52e-01**

TABLE 4.6 – Les MSE des Ratios de Treynor Généralisés

Discussion

Les limites de la variance et de la régression MCO, en présence d'outliers, permettent d'envisager d'autres outils comme le modèle MEA-Gini et le ratio de performance généralisé de Treynor-Gini. Les facteurs latents issus de l'ACP-Gini appliquée à l'univers de rentabilités d'un grand nombre de titres, seront les facteurs de risques communs pour la mise en oeuvre du modèle MEA-Gini. L'ACP-Gini a pour objectif d'expliquer les corrélations existantes entre un nombre élevé de variables par un nombre plus restreint de facteurs latents. Il s'agit d'une réduction d'espace. Cette technique d'analyse de données consiste donc à transformer des variables d'origine

en de nouvelles variables de variabilité maximale (au sens de Gini), orthogonales deux à deux et qui sont une combinaison linéaire des variables d'origine.

En présence d'outliers dans les données, l'indice de Gini capte une grande partie de la variabilité en la répartissant de manière moins concentrée sur une seule variable latente caractérisée par des outliers. Les modélisateurs ou même certains logiciels les retirent de l'échantillon avant de procéder à l'ACP classique. Le Gini étant moins sensible que la variance aux outliers, il est possible de procéder à une ACP-Gini sans être obligé de retirer les outliers du jeu de données, qui peuvent être source d'informations pertinentes, comme des rendements positifs excessifs.

4.3 Applications sur données financières

Cette partie consiste à mettre en application la théorie du modèle d'évaluation par arbitrage (MEA). Une application est faite sur le marché financier français et les composantes de l'indice français du CAC-40. Cette application se fait sur une base de données journalières de 25 actifs. D'une part, nous avons 8 actifs du marché financier français qui permettent de déterminer l'actif sans risque selon le modèle MEA. Ces actifs sont expliqués par des facteurs communs non corrélés issus de l'analyse en composantes principales de 17 facteurs. D'autre part, nous considérons diverses classes d'actifs aux caractéristiques variées pour les facteurs afin de couvrir les comportements des marchés financiers dans l'ensemble. La période considérée s'étend du 03 Janvier 2008 au 30 Décembre 2010.
Conformément aux modèles d'évaluation des actifs à l'équilibre, l'étude se fera sur les rendements des actifs. Le rendement de l'actif i sur une période $t = 1, \cdots, T$ se calcule comme suit :

$$r_{it} = \left(\frac{x_{t+1} - x_t}{x_t}\right), \quad \text{avec } i = 1, \ldots, 8 \text{ et } t = 1, \ldots, T\,; \tag{4.43}$$

où x_t représente le prix de l'action i corrigé des dividendes.

Présentation des données

Le CAC-40 est un marché financier assez connu (aspect actualité économique) et assez animé (disponibilité des données). Les actifs sélectionnés ont la particularité d'appartenir aux différents secteurs d'activités représentés sur le CAC-40. Les actions utilisés pour déterminer l'actif sans risque selon le modèle MEA sont *Danone, Legrand, L'Oréal, LVMH, Pernod Ricard, Sanofi, Sodexo et Total*.
Les facteurs dont sont issus les facteurs communs non corrélés du modèle d'évaluation d'arbitrage (MEA) sont :

— **Des indices boursiers** (*SBF 500, FTSE, DAX, Dow Jones, SP500 et Nasdaq*). Un indice boursier est une combinaison linéaire d'actions représentatives d'un marché ou d'un type précis d'entreprises.

— **Des taux de change** (le taux de change Livre sterling / Dollar américain *GBP.USD* et le taux de change Euro / Dollar américain *Euro.USD*). Le taux de change peut être défini comme le cours ou le

prix d'une devise par rapport à une autre devise.

— **Les rendements des bons du Trésor à 10 ans** : France (*fr.bond.10*), Allemagne (*deutsch.bond.10*), USA (*us.bond.10*). Un bon du Trésor est un titre d'emprunt (ou titre obligataire) émis par l'État, par l'intermédiaire du Trésor public et remboursable à échéance.

— **Des matières premières** (*Brent, Oil et Or*). Les données des matières premières sont des prix de contrats à terme portant sur l'échange de ces matières première à une date future connue à l'avance.

— **Des volatilités implicites d'indices boursiers** : DAX (*VDAX*) et SP500 (*VSP500*). La volatilité implicite représente la volatilité estimée par les acteurs d'un marché et déduite du prix des options couramment échangées sur des indices de ce marché.

— **Le taux d'inflation en France**. Le taux d'inflation correspond à la hausse ou à la baisse des prix des biens et services en terme de pourcentage sur une période donnée. L'inflation joue sur deux éléments clés de la valorisation des actions : les bénéfices et le taux d'actualisation approprié.

Analyse des données

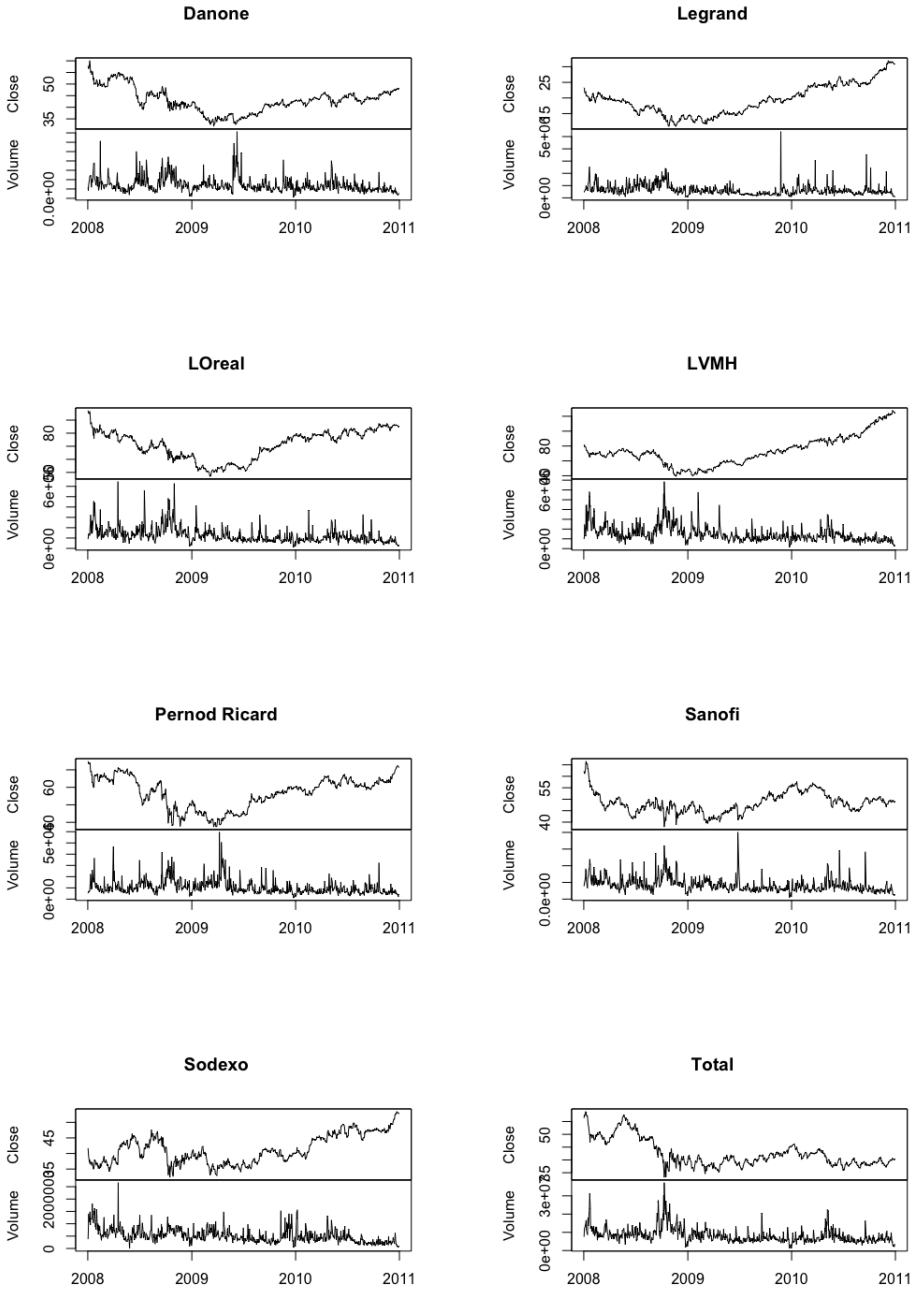


FIGURE 4.3 – Cours et Volumes échangés des actions

Les volumes d'échange sur les marchés financiers représentent les transactions entre acheteurs et vendeurs d'actifs. L'évolution d'un titre est la conséquence entre l'offre et la demande de ce dernier sur le marché, donc des volumes échangés. L'étude des volumes est par conséquent importante pour la compréhension et la prédiction de l'évolution des cours d'un titre. Les volumes peuvent confirmer la tendance du marché ou alerter sur un éventuel retournement.
Ces graphiques peuvent être segmentés en deux parties. Dans un premier temps la période 2008 à mi 2009, les actifs sélectionnés voient les cours perdre de la valeur à l'exception de Legrand, LVMH et Sodexo qui ont une période de tendance baissière moins importante (jusqu'à fin 2008 environ). En effet, début 2008 la France vient d'être atteinte par **la crise des subprimes**. Comme nous le montrent les graphiques, dans un marché baissier les volumes échangés s'intensifient jusqu'à connaitre un pique de ceux-ci, ce qui renforce la crise et provoque une perte de confiance des investisseurs pour ces valeurs. Dans un second temps, la Banque centrale européenne intervient pour une politique de stabilisation bancaire et financière. Elle rachète les dettes et injecte de l'argent dans l'économie via les banques commerciales et la facilité de prêts aux acteurs économiques en difficulté. Durant cette période, nous avons une hausse continue et croissante du cours des titres accompagnée de hausses modérées ou parfois quasi-stables des volumes échangés, à l'exception du cours et des volumes de l'action Total qui stagnent. Cela peut s'expliquer par un marché du pétrole en mauvais état. En septembre 2008, la consommation de pétrole tombe à 84 millions de barils par jour [12] en septembre 2008 et le cours du baril passe en dessous des 100 dollars. Nous sommes dans une évolution quasi-normale du marché. Les investisseurs reprennent confiance au marché.

Le tableau 4.7 résume les statistiques descriptives des rendements des actifs sélectionnés. Ce tableau met en évidence le caractère positif des rendements espérés des actifs ayant eu une période haussière beaucoup plus longue que la période baissière (Legrand, LVMH et Sodexo) et une volatilité exprimée en Gini et en écart-type relativement forte. Un paradoxe est que Sodexo a un rendement moyen positif avec un écart-type moins important que certaines actions à rendement négatifs (Pernod Ricard et Total).

12. yahoo finance

Actifs	Rendement moyen en %	Ecart-type	GMD
Danone	-0.0251	0.01920804	0.04103096
Legrand	0.0383	0.02530491	0.05465466
L'Oréal	-0.0172	0.01997087	0.04222582
LVMH	0.0556	0.02466067	0.05194678
Sanofi	-0.0315	0.02036500	0.04244031
Sodexo	0.0357	0.02048666	0.04382005
Pernod Ricard	-0.0061	0.02417532	0.04884616
Total	-0.0487	0.02219868	0.04494075

TABLE 4.7 – Statistiques descriptives des rendements des actifs sélectionnés

La volatilité des actions exprimée en GMD est quasiment égale au double de la volatilité exprimée en écart-type. Nous soupçonnons la présence d'outliers qui minimise la volatilité par la méthode de la variance et de l'écart-type. Les rendements étant compris dans l'intervalle -1 et 1, une présence d'outliers dans le calcul de la dispersion (volatilité ou risque) des rendements par une méthode quadratique (de la variance et l'écart-type) est mis à mal.

Les actifs dont sont déduits les facteurs de l'APT par la méthode statistique de l'ACP sont au nombre de 17. Pour afficher l'ensemble des facteurs, j'harmonise les variabilités de mes facteurs en divisant les valeurs des facteurs par leur écart-type (standardisation des données).

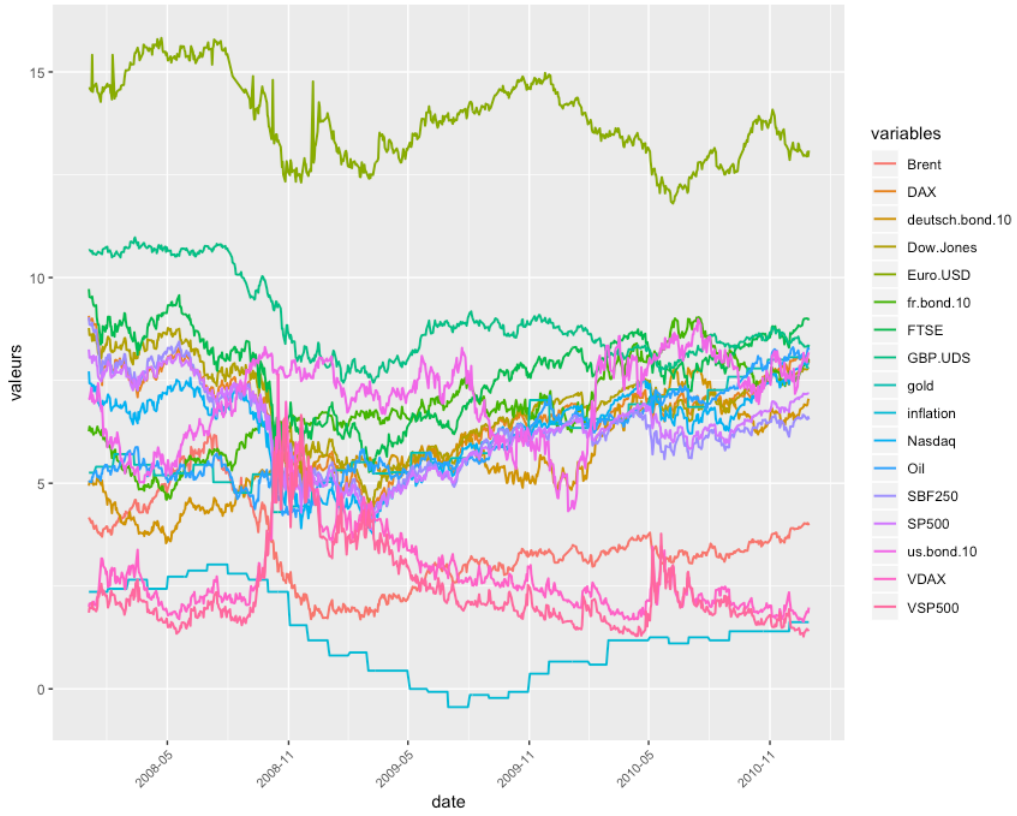


FIGURE 4.4 – Cours des facteurs standardisés

L'étude graphique des facteurs montre comme se fut le cas pour les actions, qu'ils ne sont pas stationnaires. Les indices boursiers (*SBF 500, FTSE, DAX, Dow Jones, SP500 et Nasdaq*) connaissent les mêmes tendances que les actions du CAC-40 étudiées. Dans un premier temps sur la période allant de 2008 à mi 2009, les indices boursiers ont des cours en perte de vitesse (jusqu'à fin 2008 environ) à cause de la crise des subprimes. Dans un second temps, nous avons une hausse continue et croissante de ces indices qui s'explique par l'intervention des États et des banques centrales pour la relance des économies.
Les cours du pétrole (Brent & Oil) sont en hausse sur la période avec un pique mi 2008 puis une chute sur la fin de cette année pour le Brent. Les rendements des bonds chutent fréquemment à très court terme mais ont

une tendance haussière à long terme sur la période étudiée à l'exception du rendement du bon de Trésor allemand qui grimpe sur la période 2008-2009 puis une tendance à la baisse à partir de 2009 avec un léger pique en début d'année 2010. Nous avons l'ensemble des taux de change qui fluctue constamment sur la période étudiée. Ce qui peut être révélateur de l'impact de la crise sur les économies des pays ou zone concernés (dont la France) et l'instabilité de leurs économies. La volatilité implicite du SP500 (VSP500) est très élevée du fait de la crise qui secoue les marchés à cette période et dont l'origine est les États-Unis. Cette volatilité atteint sa plus grande valeur fin 2008 en pleine crise des subprimes puis baisse jusqu'à fin 2010. La volatilité implicite du DAX (VDAX) évolue quant à elle dans le même sens que son indice. Ce qui signifie que le marché allemand s'emballe moins et est assez stable. Sur la période basse du marché, les investisseurs se sont réfugiés dans les bons du Trésor allemand, ce qui expliquerait la hausse de ces cours jusqu'en 2009.
L'inflation en France connait une hausse ininterrompue à partir de fin 2008 jusqu'à la fin de la période étudiée. Cela peut être expliquée par le quantitaive easing (ou assouplissement quantitatif) pour la relance de l'économie en baissant les coûts du crédit via les banques centrales. Avant cette période, elle connaît une baisse. Lors de la baisse des cours sur les marchés financiers, les investisseurs se réfugient dans les valeurs sûres (ou valeurs refuges) telle que l'or. La baisse du cours de l'or peut être un indicateur annonciateur de crise financière. Durant cette période, il y a une hausse des cours de l'once d'or pendant les baisses de ceux des actions. Sur l'ensemble de la période étudiée, période de crise financière, le cours est haussier.

Ces facteurs sont utilisés par la suite pour déterminer les facteurs du modèle d'évaluation par arbitrage (MEA) à travers une analyse en composantes principales (ACP).

Facteurs	Moyenne	Ecart-type	GMD
Brent ou Pétrole de la mer du nord	80.63	23.41	51.54
Oil ou Pétrole on shore	1023.74	170.59	382.69
Dow.Jones	10259.99	1488.09	3398.33
Euro.USD	1.40	0.10	0.23
FTSE	5131.63	666.37	1506.06
GBP.USD	1.66	0.19	0.40
DAX	5803.25	871.83	1976.67
SP500	1102.11	175.36	401.26
Nasdaq	2116.67	337.11	745.59
SBF250	2713.74	432.17	971.71
VDAX	26.78	9.65	19.42
VSP500	29.02	12.14	24.10
deutsch.bond.10	3.35	0.60	1.35
fr.bond.10	3.67	0.54	1.21
us.bond.10	3.37	0.48	1.07
Or	33001.04	5479.63	12317.35
inflation	1.74	1.36	3.09

TABLE 4.8 – Statistiques descriptives des facteurs

Rappelons que l'écart-type étant la racine carrée de la variance, la dispersion mesurée par la variance est beaucoup plus importante pour la plupart des facteurs retenus.

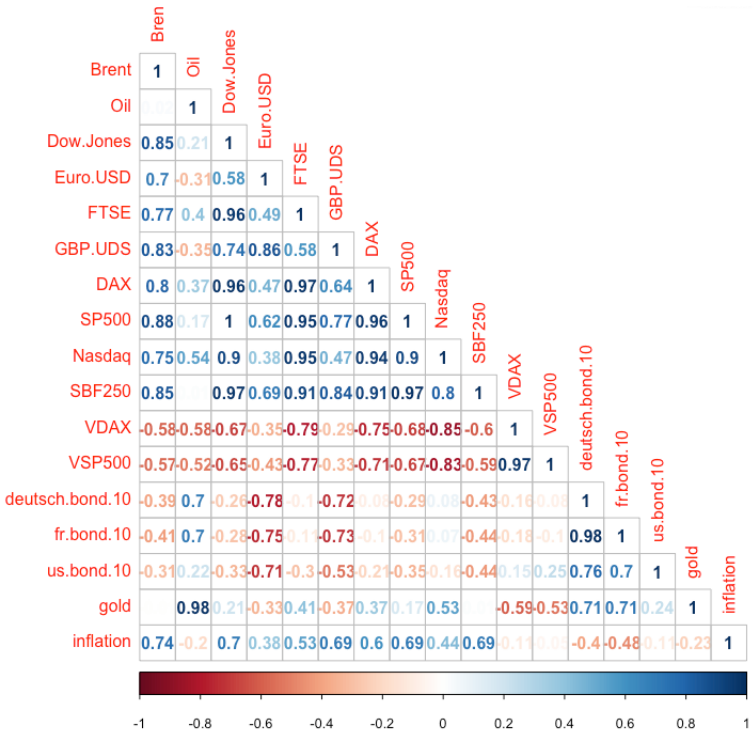


FIGURE 4.5 – Corrélation de Pearson des facteurs

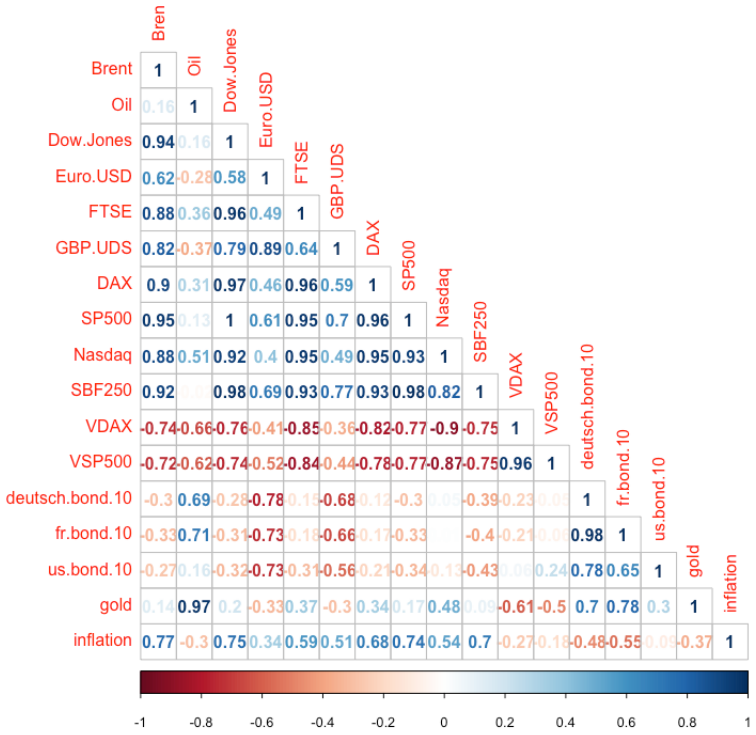


FIGURE 4.6 – Corrélation au sens de la GMD des facteurs

Résultats du Modèle d'Evaluation par Arbitrage

La méthodologie utilisée est un processus à deux étapes. La première étape consiste à faire, à partir du cours des 17 facteurs retenus, une analyse en composantes principales d'une part, par la méthode classique (ACP classique) et d'autre part par la méthode du Gini (ACP-Gini). De l'ACP, il est déterminé les facteurs du modèle d'évaluation par arbitrage(MEA). Les facteurs du MEA correspondent aux axes factoriels retenus à l'issue de l'ACP. Dans un premier temps, les outils d'aide à l'interprétation tels que la projection des variables et les valeurs propres permettront de sélectionner les axes les plus significatifs. Puis dans un second temps, la corrélation des variables avec les composantes principales (CP) retenues et la contribution des variables à la formation des CP permettront d'interpréter les primes de risque.
La deuxième étape consiste à faire une régression multiple (régression par

la méthode des moindres carrés ordinaires (MCO) et une régression Gini) du rendement des actifs sur le rendement des facteurs afin de déterminer la valeur de l'actif sans risque et des primes de risque à partir d'une combinaison des sensibilités du rendement des actifs aux rendements des facteurs de risque retenus.

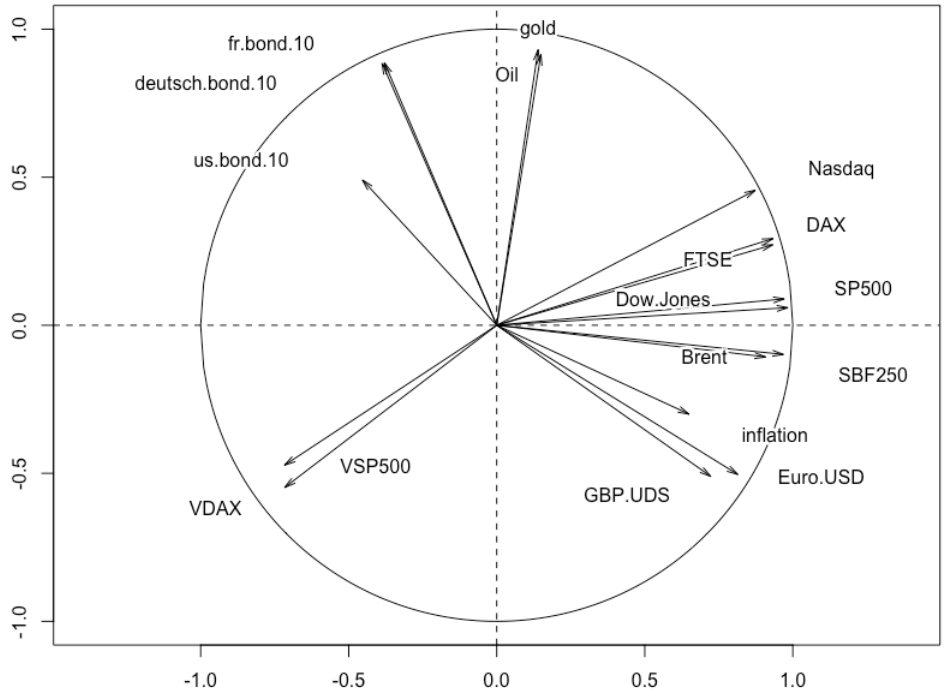


FIGURE 4.7 – Projection classique des variables

Le graphique illustre la projection des variables sur les deux premiers axes factoriels. Les deux axes les plus importants en terme de pourcentage de variabilité expliquée. L'axe factoriel 1 oppose les indices boursiers, le Brent et les taux de change aux volatilités des indices boursiers du SP500 (VSP500) et du DAX (VDAX). L'axe 2 est principalement formé de l'or, du pétrol on shore (Oil ou WTI), des bons du Tésor français et allemand à 10 ans.

Résultats du MEA classique

A partir de la représentation graphique des valeurs propres et du pourcentage cumulé de variance expliquée, l'on détermine les axes factoriels retenus à l'issue de l'analyse en composantes principales.

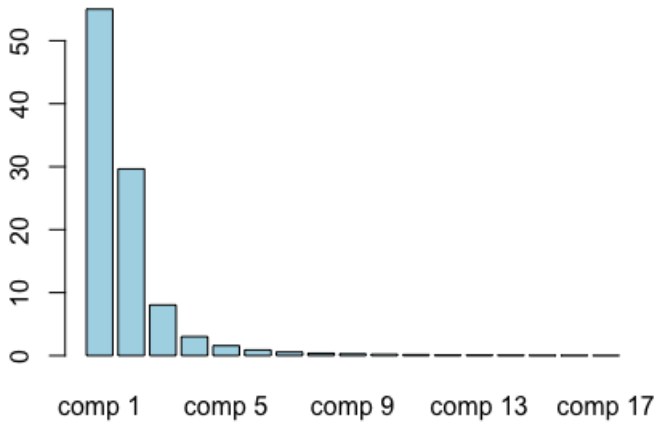


FIGURE 4.8 – Valeurs propres - ACP classique

Par la technique du coude, les deux premiers axes peuvent être retenus comme les axes factoriels pertinents. Il existe une rupture au delà de la seconde composante principale. La variance cumulée, exprimée en pourcentage permet d'appuyer la représentation graphique.

Composantes principales	**Pourcentage cumulée** (%)
Composante 1	55.01
Composante 2	84.66
Composante 3	92.7

TABLE 4.9 – Pourcentage de variance cumulée

En effet, les deux premières composantes (ou axes factoriels) retenues représentent à elles 84.7% de la variance expliquée. Ces deux axes contiennent donc environ 85% de l'information contenue dans la matrice des 17 facteurs de départ.

Facteurs	Dimension 1	Dimension 2
Brent ou Pétrole de la mer du nord	0.908	-0.107
Oil ou Pétrole on shore	0.149	0.914
Dow.Jones	0.971	0.089
Euro.USD	0.722	-0.510
FTSE	0.934	0.293
GBP.USD	0.814	-0.504
DAX	0.932	0.272
SP500	0.983	0.059
Nasdaq	0.874	0.456
SBF250	0.968	-0.098
VDAX	0.716	-0.548
VSP500	-0.717	-0.473
deutsch.bond.10	-0.379	0.886
fr.bond.10	-0.387	0.885
us.bond.10	-0.454	0.491
Or	0.140	0.931
inflation	0.649	-0.300

TABLE 4.10 – Corrélation des variables avec les axes 1 & 2

Les variables qui caractérisent le plus l'axe 1 selon l'ACP classique sont le Brent, le Dow Jones, le FTSE, le taux de change GBP.USD, le DAX, le SP500, le Nasdaq, le SBF250. Cette composante, étant une composante linéaire pour la plupart des indices boursiers, pourrait être assimilé à un indice boursier international. Quant à la deuxième composante principale qui est caractérisée par l'or, les rendements des bonds du trésor allemand (deutsch.bond.10) et français (fr.bond.10) à 10 ans et du pétrole one shore (Oil) peut être assimilée à un facteur de valeurs refuges.
La contribution absolue (CTA) permettent de confirmer les variables qui caractérisent les axes factoriels retenus. De plus, les CTA donnent des

informations sur la part de risque rémunérée pour chacune des variables dans la prime de risque assimilée aux différents facteurs de risque (axes factoriels).

Facteurs	CTA 1	CTA 2
Brent ou Pétrole de la mer du nord	8.8192187	0.22747662
Oil ou Pétrole on shore	0.2376149	16.59253044
Dow.Jones	10.0881583	0.15691382
Euro.USD	5.5799521	5.16557609
FTSE	9.3181828	1.70095746
GBP.USD	7.0930663	5.04391001
DAX	9.2940716	1.46658075
SP500	10.3410564	0.06967172
Nasdaq	8.1662506	4.13211192
SBF250	10.0266713	0.19252359
VDAX	5.4855940	5.95315284
VSP500	5.4929457	4.43212927
deutsch.bond.10	1.5356291	15.58042419
fr.bond.10	1.6036376	15.52850202
us.bond.10	2.2003815	4.77910547
Or	0.2090162	17.19120219
inflation	4.5085529	1.78723161

TABLE 4.11 – Contribution absolue des variables à la formation des axes 1 & 2

Les variables retenues sont celles ayant une contribution absolue supérieure à 1/nombre de facteurs x 100 (1/17x100 = 5.88%). Les variables retenues plus haut comme celles contribuant à la formation des axes 1 et 2 sont correctes. De plus, les CTA donnent des informations sur la part de risque rémunérée pour chacune des variables dans la prime de risque assimilée aux différents facteurs de risque (axes factoriels). Nous aurons par exemple 8.81% de la prime de risque liée au premier facteur de risque qui est dû à la présence du Brent dans ce facteur.

FIGURE 4.9 – Composantes principales

La première composante principale pourrait être interprétée comme un indice boursier international. Il est observé une chute brutale des cours de cet indice entre janvier 2008 et environ février ou mars 2009 puis une augmentation du cours sur le reste de la période avec une légère chute en juin 2010. La seconde courbe est celle de la seconde composante principale ou le facteur de valeurs refuges. Il connait une légère chute en janvier et environ mars 2008. Ce qui est un signal pour la crise de 2008. Le cours de ce facteur connait une légère hausse puis une légère chute aux environs de décembre 2008. Cette chute coincide à celle de la première composante. Cela peut être dû à un manque de confiance aux marchés de la part des investisseurs. Ensuite les cours de la composante à valeurs refuges connaît une hausse durant le reste de la période. Les investisseurs jouent la carte de la sécurité, ils investissent dans les valeurs refuges. Face à une augmen-

tation de la demande des valeurs refuges, les cours continuent à prendre de la valeur.

A cette étape, une régression multiple des rendements sur les rendements des facteurs obtenus est effectuée afin de récupérer les sensibilités estimées (bêta $\hat{\beta}$) issues de cette régression,

$$r_{it} = \beta_{i1}\, F_1 + \beta_{i2}\, F_2 \, + \boldsymbol{\epsilon}_{i,t} \quad \text{avec } i = 1, \cdots, 8.$$

Actifs	β**.Facteur 1**	**T-stat**	β**.Facteur 2**	**T-stat**
Danone	0.8160657	18.129	-0.1013410	-1.939
Legrand	1.1048905	18.727	-0.1759275	-2.568
L'Oréal	0.937585	20.906	-0.143218	-2.751
LVMH	1.3479264	27.958	-0.1823257	-3.258
Sanofi	0.8234931	16.783	-0.1073978	-1.885
Sodexo	0.8843652	18.474	-0.1251178	-2.251
Pernod Ricard	1.0299718	18.232	-0.1385495	-2.113
Total	1.19234035	27.554	-0.08951544	-1.782

TABLE 4.12 – Les bêtas issus de la régression multiple MCO sans constante

Nous avons des sensibilités positives avec les rendements du premier facteur puis négatives avec les rendements du deuxième facteur. Les premières sensibilités sont assez élevées, certaines supérieures à 1 tandis que les secondes sensibilités sont relativement faibles, proches de zéro. Avec un risque d'erreur $\alpha = 5\%$, les estimateurs significativement différents de zéro sont marqués en bleu.
En présence d'erreurs de mesures (ce qui n'est pas le cas) ou d'outliers dans les données des régresseurs (ou variables explicatives), il est démontré que les estimateurs par les MCO sont biaisés et non convergents (Voir Chapitre 1). Un test de détection d'une la présence éventuelle d'outliers dans les données des deux variables explicatives est effectué. Une représentation graphique par la boîte à moustache (ou boxplot) est effectuée afin de déterminer les observations des rendements des facteurs susceptibles d'être des outliers.

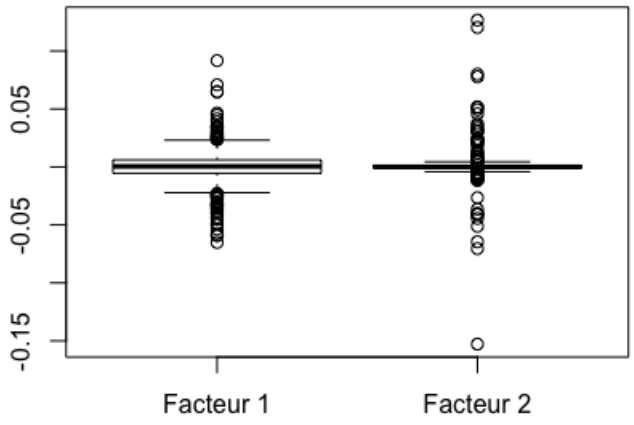


FIGURE 4.10 – Boîte à moustache des rendements des facteurs retenus

Graphiquement, il est soupçonné l'existence de valeurs atypiques (ou outliers) dans les rendements de ces facteurs. Un test de Rosner est effectué pour vérifier l'exactitude des soupçons de la présence d'outliers. Le test de Rosner se fait comme suit :

1. Les n observations sont ordonnées de la plus petite valeur à la plus grande.
2. Le nombre maximum d'observations aberrantes k suspectées est spécifié, où k est compris entre 1 et 10.
3. Une série de tests statistiques est effectuée. L'observation (grande ou petite) la plus éloignée de la moyenne est supprimée et en recomposant la statistique de test selon l'équation suivante :

$$R_{i+1} = \frac{|x^{(i)} - \bar{x}^{(i)}|}{s^{(i)}},$$

où $\bar{x}^{(i)}$ et $s^{(i)}$ sont respectivement la moyenne et l'écart-type des données après que les observations i aient été retirées ;

$x^{(i)}$ est l'observation dans le sous-ensemble de la base de données la plus éloignée de $\bar{x}^{(i)}$.

Une fois tous les tests statistiques ($R_1, \cdots R_k$) calculés, une série de tests est effectuée. Il est testé d'abord l'hypothèse qu'il existe k outliers en comparant R_k à la valeur critique λ_k, obtenue d'une table (Table A-4, EPA) pour le niveau de significativité spécifié α.

Décision :

$R_k > \lambda_k$, alors le test est significatif et nous pouvons rejeter l'hypothèse nulle qu'il n'y a pas d'outliers dans les données et conclure que les k valeurs extrêmes sont des outliers.

$R_k \leq \lambda_k$, nous allons tester l'hypothèse selon laquelle $k-1$ outliers en comparant R_{k-1} à la valeur critique λ_{k-1}. Ce processus se poursuit jusqu'à ce que l'un des tests soit significatif et que nous puissions conclure qu'il y a un certain nombre d'outliers dans les données, ou jusqu'à ce que tous les tests aient été effectués et qu'aucun n'ait été jugé significatif. Si aucun des tests n'est significatif, nous pouvons conclure alors qu'il n'y a pas d'outliers dans les données.

Données	**Taille de l'échantillon**
Les rendements du facteur 1	745
Paramètres des tests statistiques " k "	**Niveau de risque " α "**
3	5% ou 0.05

i	**Num. Obs**	**Test statistique (R_{i+1})**	**Valeur critique (λ_{i+1})**	**Outliers**
0	196	6.451671	3.966394	Vrai
1	207	5.136207	3.966053	Vrai
2	235	4.791061	3.965712	Vrai

TABLE 4.13 – Test de Rosner pour les rendements du facteur 1

Données	Taille de l'échantillon
Les rendements du facteur 2	745
Paramètres des tests statistiques " k "	**Niveau de risque " α "**
3	5% ou 0.05

i	Num. Obs	Test statistique (R_{i+1})	Valeur critique (λ_{i+1})	Outliers
0	193	12.53866	3.966394	Vrai
1	458	11.58712	3.966053	Vrai
2	235	12.16281	3.965712	Vrai

TABLE 4.14 – Test de Rosner pour les rendements du facteur 2

Le test de Rosner pour un niveau de risque $\alpha = 5\%$ et un test de détection simultanée de trois outliers (3) confirme que les trois plus grandes valeurs en valeur absolue sont des outliers comme l'indique les tableaux. Par conséquent, les bêtas estimés ne sont pas fiables.
A partir des bêtas (relatifs aux facteurs 1 & 2) des actifs issus de la régression par les MCO et des hypothèses d'absence d'arbitrage $\sum_i^8 w_i = 1$ et $\sum_i^8 \ w_i\widehat{\beta}_i = 0$, où w_i représente le poids des actifs considérés (Danone, Legrand, L'Oréal, etc.) dans la composition de l'actif sans risque, alors $\sum_i^8 \ w_i\mathbb{E}(R_i) \approx rf$ où R_i est le rendement de l'actif i et rf l'actif sans risque. Nous avons les pondérations de l'actif sans risque grâce au système d'équations suivant :

$$\begin{aligned} w_1 + w_2 + \cdots + w_8 &= 1 \\ \widehat{\beta}_{1,1}w_1 + \widehat{\beta}_{2,1}w_2 + \cdots + \widehat{\beta}_{8,1}w_8 &= 0 \quad (\beta_{i,1} : \text{sensibilités au facteur 1}) \\ \widehat{\beta}_{1,2}w_1 + \widehat{\beta}_{2,2}w_2 + \cdots + \widehat{\beta}_{8,2}w_8 &= 0 \quad (\beta_{i,2} : \text{sensibilités au facteur 2}) \end{aligned}$$

Le Tableau 4.15 résume le poids des 8 actifs considérés constituant l'actif sans risque dans le modèle d'évaluation par arbitrage classique.

Actifs	Pondération
Danone	1.02
Legrand	-0.44
L'Oréal	0.34
LVMH	-1.33
Sanofi	0.95
Sodexo	0.63
Pernod Ricard	0.05
Total	-0.22

TABLE 4.15 – Le poids des actifs constituant l'actif sans risque

L'actif sans risque devrait donc être égale selon ce modèle et les actifs considérés à **-0.12 %**. Nous avons un actif sans risque qui a un rendement négatif. Cela est dû à la période considérée, qui est la période de la crise des subprimes. L'actif sans risque est rémunéré négativement mais reste une valeur sûre avec une volatilité nulle durant cette période d'incertitude. Il est à présent nécessaire de déterminer les primes de risque liées à chaque facteur de risque de notre modèle APT. La prime de risque est l'excès de rendement entre le rendement attendu du facteur de risque et l'actif sans risque.

Les rendements attendus sont pour le facteur 1 (**2.86e-05**) avec une prime de risque unitaire (**1.22e-03**) et pour le facteur 2 (**5.17e-04**) avec une prime de risque unitaire (**1.71e-03**). L'espérance de rendement des actifs est calculée selon le modèle d'évaluation par arbitrage (MEA).

Actifs	Moyenne des rendements	Moyenne des rendements MEA
Danone	-6.74e-05	-3.7e-04
Legrand	7.03e-04	-1.44e-04
L'Oréal	2.71e-05	-2.93e-04
LVMH	8.6e-04	1.42e-04
Sanofi	-1.08e-04	-3.71e-04
Sodexo	5.67e-04	-3.27e-04
Pernod Ricard	2.3e-04	-1.72e-04
Total	-2.4e-04	1.11e-04

TABLE 4.16 – L'espérance de rendement selon le modèle MEA

Force est de constater que les actifs sont mal évalués. Il existe cependant un écart assez considérable des espérances de rendement entre ces deux méthodes.

Résultats du MEA-Gini

A partir de la représentation graphique des valeurs propres et du pourcentage cumulé de GMD expliqué, l'on détermine les axes factoriels retenus à l'issue de l'analyse en composantes principales par une approche Gini.

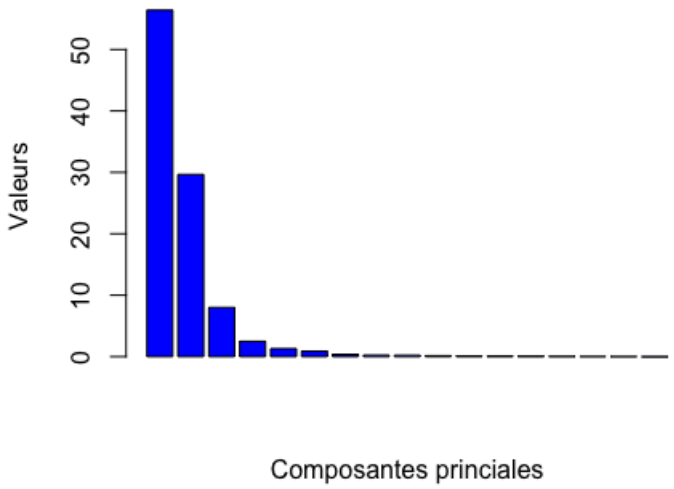


FIGURE 4.11 – Valeurs propres - ACP Gini

Tout comme dans le cas de l'ACP classique, la technique du coude est utilisée. Les deux premiers axes sont retenus comme les axes factoriels pertinents. Il existe une rupture au delà de la seconde composante principale. Le GMD cumulé, exprimé en pourcentage permet d'appuyer la représentation graphique.

Composantes principales	**Pourcentage cumulée** (%)
Composante 1	56.41
Composante 2	86.07
Composante 3	94.07

TABLE 4.17 – Pourcentage de variance cumulée

En effet, les deux premières composantes (ou axes factoriels) retenues représentent à elles 86.07% de la GMD expliqué. Ces deux axes contiennent 86% de l'information contenue dans la matrice des 17 facteurs de départ.

La corrélation des variables avec les axes factoriels dans le cas de l'ACP-Gini nous permet de connaître les variables qui participent fortement à la formation des axes.

Facteurs	Dimension 1	Dimension 2
Brent ou Pétrole de la mer du nord	-0.949	0.024
Oil ou Pétrole on shore	-0.135	0.906
Dow.Jones	-0.974	0.068
Euro.USD	-0.705	-0.502
FTSE	-0.935	0.233
GBP.USD	-0.755	-0.450
DAX	-0.937	0.224
SP500	-0.983	0.052
Nasdaq	-0.866	0.405
SBF250	-0.980	-0.034
VDAX	0.717	-0.578
VSP500	0.725	-0.436
deutsch.bond.10	0.422	0.879
fr.bond.10	0.337	0.926
us.bond.10	0.392	0.559
Or	-0.092	0.932
inflation	-0.635	-0.225

TABLE 4.18 – Corrélation des variables avec les axes 1 & 2 au sens du Gini

Tout comme dans le cas de l'ACP classique, les variables qui caractérisent le plus l'axe 1 selon l'ACP Gini sont le Brent, le Dow Jones, le FTSE, le taux de change GBP.USD, le DAX, le SP500, le Nasdaq, le SBF250. Cette composante étant une composante linéaire qui pourrait être assimilée à un indice boursier international. Quant à la deuxième composante principale qui est caractérisée par l'or, les rendements des bonds du trésor allemand (deutsch.bond.10) et français (fr.bond.10) à 10 ans et du pétrole one shore (Oil) peut être assimilée à un facteur de valeurs refuges. Il existe cependant quelque différence au niveau des signes.

Les deux premiers facteurs étant retenus, une régression Gini multiple des rendements des actifs sur les rendements des facteurs obtenus est

effectuée afin de récupérer les sensibilités estimées ($\hat{\beta}^G$) issues de cette régression

$$r_{it} = \beta_{i1}^G F_1 + \beta_{i2}^G F_2 + \boldsymbol{\epsilon}_{i,t} \quad \text{avec } i = 1, \cdots, 8.$$

Avant d'aller plus loin, un test de significativité des paramètres est effectué pour s'assurer de la validité du modèle retenu. Les estimateurs étant des U-statistiques, l'écart-type estimé de l'estimateur $\widehat{\sigma}_\beta$ peut être obtenu par la méthode du Jackknife. Pour cela, il est nécessaire de calculer la variance estimée puis l'écart-type estimé par la méthode de Jackknife. De manière générale, l'estimateur de la variance par Jackknife d'un estimateur U est donné par :

$$\hat{V}(\hat{\beta}^G) = \frac{n-1}{n} \sum_{i=1}^{n} \left[U_{-i} - \frac{1}{n} \sum_{i=1}^{n} U_{-i} \right]^2$$

Pour tester une hypothèse $H_0 : \beta^G = 0$, nous allons comparer l'estimateur normalisé

$$Z = \frac{\sqrt{n}\,(\hat{\beta}^G - 0)}{\sqrt{\hat{V}(\hat{\beta}^G)}},$$

où n est le nombre d'observations et Z qui suit une loi normale avec un niveau de confiance donné α (95 % en général).

Actifs	β**.Facteur 1**	**T-stat**	β**.Facteur 2**	**T-stat**
Danone	0.8450847	3.596224	-0.2145979	-3.094844
Legrand	1.1164029	4.074794	0.1710407	1.450870
L'Oréal	1.0676923	3.538923	-0.4108431	-2.613698
LVMH	1.41378022	4.1258507	0.05184407	0.5236806
Sanofi	0.8842522	3.478800	-0.2669564	-3.609115
Sodexo	0.8536561	5.586959	0.1689376	1.157014
Pernod Ricard	0.9841737	3.837096	-0.1952901	-1.376892
Total	0.9819558	3.658786	0.2472265	1.324645

TABLE 4.19 – Les bêtas issus de la régression multiple Gini sans constante

Nous avons des sensibilités positives avec les rendements du premier facteur tout comme dans le cas de la régression par les MCO avec quelques légères différences. Ces actifs ont des sensibilités proches de l'unité. Elles suivent la tendance de l'indice international avec certains actifs offensifs tels que LVMH, Legrand et L'Oréal. Contrairement à la régression par les MCO, les sensibilités des rendements des actifs sont positives à l'exception de Danone, L'Oréal, Sanofi et Pernod Ricard qui sont négatives. Ces actions restent défensives avec les rendements du deuxième facteur. Avec un risque $\alpha = 5\%$ d'erreur, les estimateurs significativement différents de zéro sont marqués en bleu.
Comme dans le cas du MEA classique, à partir des bêtas (relatifs aux facteurs $1 \,\&\, 2$) des actifs, cette fois issus de la régression Gini et des hypothèses d'absence d'arbitrage, nous avons les pondérations de l'actif sans risque grâce au système d'équations suivant :

$$\begin{aligned} w_1 + w_2 + \cdots + w_8 &= 1 \\ \widehat{\beta}^G_{1,1} w_1 + \widehat{\beta}^G_{2,1} w_2 + \cdots + \widehat{\beta}^G_{8,1} w_8 &= 0 \quad (\beta^G_{i,1} : \text{sensibilités au facteur 1}) \\ \widehat{\beta}^G_{1,2} w_1 + \widehat{\beta}^G_{2,2} w_2 + \cdots + \widehat{\beta}^G_{8,2} w_8 &= 0 \quad (\beta^G_{i,2} : \text{sensibilités au facteur 2}) \end{aligned}$$

Le Tableau 4.20 résume le poids des 8 actifs considérés constituant l'actif sans risque dans le MEA-Gini.

Actifs	**Pondération**
Danone	0.75
Legrand	- 0.13
L'Oréal	-0.35
LVMH	-1.5
Sanofi	0.55
Sodexo	1.00
Pernod Ricard	0.17
Total	0.51

TABLE 4.20 – Le poids des actifs constituant l'actif sans risque

Pareillement que par la méthode classique du MEA, l'actif sans risque

est négatif. Selon ce modèle et les actifs considérés, l'actif sans risque devrait donc être égale à **-1.017e-03**. Cela est dû à la période considérée, la période de la crise des subprimes. L'actif sans risque est rémunéré négativement mais reste une valeur sûre avec une volatilité nulle durant cette période d'incertitude. Il est à présent nécessaire de déterminer les primes de risque liées à chaque facteur de risque de notre modèle APT (ou MEA) par la méthode du Gini. La prime de risque est l'excès de rendement entre le rendement attendu du facteur de risque et l'actif sans risque. Le rendement moyen ou l'espérance de rendement de chaque actif est calculé, puis l'on retranche la valeur de l'actif sans risque.

Les rendements attendus pour le facteur 1 sont (**2.34e-05**) avec une prime de risque unitaire (**1.04e-03**) et pour le facteur 2 (**5.34e-04**) avec une prime de risque unitaire (**1.55e-03**). L'espérance de rendement des actifs est calculée selon le modèle d'évaluation par arbitrage en Gini (MEA-Gini).

Actifs	**Moyenne des rendements**	**Moyenne des rendements MEA-Gini**
Danone	-6.74e-05	-4.7e-04
Legrand	7.03e-04	4.1e-04
L'Oréal	2.72e-05	-5.43e-04
LVMH	8.6e-04	5.34e-04
Sanofi	-1.1e-04	-5.11e-04
Sodexo	5.67e-04	1.33e-04
Pernod Ricard	2.3e-04	-2.96e-04
Total	-2.4e-04	3.88e-04

TABLE 4.21 – L'espérance de rendement selon le modèle MEA-Gini

Tout comme dans le modèle classique du MEA, les actifs sont mal évalués. A l'exception de l'action TOTAL qui est sous-évaluée, les autres actions sont sur-évaluées. Il existe un écart comme cela est le cas pour l'estimation de l'espérance de rendement des actifs avec méthode classique du MEA.

Actifs	Moyenne arithmétique	MEA-Gini	MEA classique
Danone	-6.74e-05	-4.7e-04	-3.7e-04
Legrand	7.03e-04	4.1e-04	-1.44e-04
L'Oréal	2.72e-05	-5.43e-04	-2.93e-04
LVMH	8.6e-04	5.34e-04	1.42e-04
Sanofi	-1.1e-04	-5.11e-04	-3.71e-04
Sodexo	5.67e-04	1.33e-04	-3.27e-04
Pernod Ricard	2.3e-04	-2.96e-04	-1.72e-04
Total	-2.4e-04	3.88e-04	1.11e-04

TABLE 4.22 – La comparaison des espérances de rendement des différents actifs

Résultats du ratio de Treynor généralisé avec une approche Gini

Le ratio de performance de Treynor généralisé est un ratio de performance qui s'appuie sur un modèle d'évaluation multi-factoriel (ou multi-indice) tel que le modèle de Fama et French, le modèle Carhart ou le modèle d'évaluation par arbitrage via l'analyse en composantes principales. L'objectif est de savoir si notre stratégie permet de battre un portefeuille de référence (Benchmark). Travaillant avec des actifs du marché financier, le portefeuille de référence choisi est un indice du marché français, en l'occurrence le CAC 40. Le choix se porte sur le CAC 40 qui reste l'indice phare des 40 valeurs mobilières les plus performantes du marché français et les actions qui constituent notre portefeuille appartiennent au CAC 40.
Le calcul du ratio généralisé de Treynor se fait dans cette application à partir du modèle d'évaluation par arbitrage à facteurs latents en Gini. Le ratio de performance généralisé de Treynor pour un portefeuille P avec une approche Gini s'écrit pour rappel comme suit :

$$RTG_p^G = (\mathbb{E}(r_p) - r_f) \cdot \frac{\sum_{j=1}^{k} \tilde{\mu}_j \beta_{mj}}{\sum_{j=1}^{k} \frac{\beta_{pj}^G}{\beta_{mj}^G} \tilde{\mu}_j \beta_{mj}}$$

$$RTG_p^G = (\mathbb{E}(r_p) - r_f) \cdot \frac{\sum_{j=1}^{k} w_j}{\sum_{j=1}^{k} \beta_{pj}^{G*} w_j}$$

où $j = 1, \cdots, k$ désigne le nombre de primes de risque distinctes, $w_j = \tilde{\mu}_j \beta_{mj}^G$ avec $\beta_{mj}^G = \beta_{m1}^G, \cdots, \beta_{mk}^G$ les sensibilités du portefeuille de référence aux différents risques systématiques avec $\tilde{\mu}_j$ la prime de risque associée à chaque facteur et $\beta_{pj}^{G*} = \frac{\beta_{pj}^G}{\beta_{mj}^G}$ avec $\beta_{pj}^G = \beta_{p1}^G, \cdots, \beta_{pk}^G$ les sensibilités du portefeuille P aux différents risques systématiques. Il est aussi possible de le calculer pour un actif i . Dans l'application que nous avons il est possible de calculer le ratio de Treynor généralisé pour chaque actif et en faire un classement en fonction de leur performance.
L'application se fait sur un modèle APT-Gini basé sur l'analyse en composantes principales avec l'approche Gini. A partir des 8 actions [13] du marché financier français que nous avons, nous construisons un portefeuille efficient de même espérance de rentabilité que le Benchmark. L'objectif étant de voir si à niveau égal de rendement à long terme, notre portefeuille bat le benchmark.
En nous appuyant sur les résultats du *Théorème 2.2*, le vecteur $\mathbf{p} = (p_1, \ldots, p_8)$, issu des 8 premières valeurs de la matrice $\mathbf{z}_p$ est pour un rendement cible α donné et une matrice Gini-Cogini **G** inversible, le vecteur des poids des actifs qui composent le portefeuille efficient de rendement α. Le benchmark a pour rendement moyen **-0.03**%. Le poids des actifs constituants un portefeuille efficient d'espérance rendement $\alpha =$ **-0.03**% par une approche Gini sont :

13. DANONE, LEGRAND, L'OREAL, LVMH, SANOFI, SODEXO, PERNOD RICARD & TOTAL.

Actifs	Pondération en %
Danone	3.13e-03
Legrand	7.29e+01
L'Oréal	2.01e+00
LVMH	-2.17e+02
Sanofi	1.09e+02
Sodexo	2.27e+02
Pernod Ricard	-6.98e+01
Total	-2.44e+01

TABLE 4.23 – Le poids des actifs

Le portefeuille dont la performance est évaluée est constitué comme représenté dans le tableau ci-dessus. Afin d'obtenir les sensibilités aux différents risques systématiques (déjà connus) du portefeuille de référence et du portefeuille constitué, deux régressions multiples par une approche Gini sont réalisées.

	Valeurs	T.stat
$\widehat{\beta}^G_{m1}$	1.22e-03	10.77114
$\widehat{\beta}^G_{m2}$	-6.44e-02	-0.5657058

TABLE 4.24 – Les estimateurs du Benchmark

Seul le $\widehat{\beta}^G_{m1}$ est significativement différent de zéro pour un seuil de 5%.

	Valeurs	T.stat
$\widehat{\beta}^G_{p1}$	-2.7e-04	-3.084855
$\widehat{\beta}^G_{p2}$	2.65e-04	5.038609

TABLE 4.25 – Les estimateurs du Portefeuille

Les $\widehat{\beta}^G_{pj}$ sont significativement différents de zéro pour un seuil de 5%.

	Valeurs
$\widehat{\beta}_{p1}^{G*}$	-2.12e-01
$\widehat{\beta}_{p2}^{G*}$	-2.66
Prime liée au Facteur 1	1.03e-03
Prime liée au Facteur 2	1.55e-03
$\sum_{j=1}^{2} w_j$	1.17e-03
$\sum_{j=1}^{2} \beta_{pj}^{G*} w_j$	-4.73e-06
Prime du Portefeuille	7.24e-04
RTG_p^G	**-1.79e-01**

TABLE 4.26 – Le ratio généralisé de Treynor - Gini

Le RTG_p^G est de **-1.79e-01**, nous avons une rentabilité très peu intéressante. La prime de risque par unité de risque pondérée est négative. La stratégie idéale est de détenir le portefeuille de marché comme stratégie.

4.4 Conclusion

Dans ce chapitre, nous avons montré par des Théorèmes qu'il est possible d'appliquer à ses méthodes de la Finance une approche Gini. Face à la méthode classique (approche par la variance du MEA et du RTG), les simulations de Monte Carlo illustrent la robustesse des primes de risque en présence de valeurs extrêmes issues du MEA-Gini, des sensibilités Gini (voir Chapitre 1) et donc du MEA-Gini. De même avec des très faibles Mean Square Error (MSE), la stabilité du ratio de Treynor généralisé par l'approche Gini est mise en avant. De plus, les simulations de Monte Carlo sur les RTG montrent que l'approche Gini du RTG par des MSE plus faibles que ceux de l'approche classique est moins sensible à la présence d'outliers dans les données.
Une application sur le modèle d'évaluation par arbitrage et sur le ratio de Treynor généralisé est faite. Elle est réalisée sur des données financières de 8 actifs du marché financier français et d'un certain nombre de facteurs communs.

BIBLIOGRAPHIE

Abhijit Banerjee. (2010). *A multidimensional Gini index*. Mathematical Social Sciences, **vol. 60**, pages 87 à 93.

Arthur Charpentier., Stephane Mussard et Téa Ouraga. (2020). *Principal Component Analysis : A Generalized Gini Approach*. European Journal of Operational Research, in under revew.
. (2008)]AD08 Aswath Damodaran. (2008). *Equity Risk Premiums (ERP) : Determinants, estimation and implication ?*. www.damodaran.com.

Eugene Fama., Kenneth French. (1992). *The Cross-Section of Expected Stocks Returns*. The Journal of Finance, **vol. XLVII**, n. 2.

Eugene Fama., Kenneth French. (2015). *A five-factor asset pricing model*. Journal of Financial Economics, **vol. 116**, n. 1, pages 1 à 22.

Eugene Fama., Kenneth French. (2016). *Dissecting anomalies with a five-factor model* . Review of Financial studies, **vol. 29**, n. 1, pages 69 à 103.

Florin Aftalion. (2004). *Le MEDAF après quarante ans*. Banques et Marchés n. 73, pages 56 à 62.

George Hübner. (2003). *The Generalized Treynor Ratio : A note*. Université de Liège, Management Working Paper

Haim Shalit and Sholmo Yitzhaki. (1984). *Mean-Gini, portfolio Theory, and the Pricing of Risy Assets*. The Journal of Finance. Université de Liège, **vol. XXXIX**, n. 5.

Harry Markowitz (1952). *Portfolio selection*. Journal of Finance, **vol. 7**, n. 1, pages 77 à 91.

Harry Markowitz (2005). *Market efficiency : a theoretical distinction and so what ?*. Financial Analysts Journal, **vol. 61**, n. 5, pages 17 à 30.

Jan Mossin. (1966). *Equilibrium in a Capital Asset market*. Econometrica, **vol. 34**, n. 4, pages 768 à 783.

John Lintner. (1965). *The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets*. Review of Economics and Statistics, **vol. 47**, n. 1, pages 13 à 37.

Luc Dumontier.(2016). *Les fondements académiques des solutions d'investissement factoriel*. Revue Banque, n. 798, pages 70 à 74.

Mark Carhart. (1997). *On persistence in Mutual Fund Performance*. The Journal of Finance, **vol. 52**, n. 1, pages 57 à 82.

Massoud Mussavian. (1998). *Evaluation des risques : l'alternative du modèle APT*. L'Art de la Finance, Les Echos.

Moussa Mbairadjim., Kamdem Sadefo. (2016). *A fuzzy multifactor asset pricing model*. The Journal of Finance, Researchgate.

Richard Roll. (1977). *A critique of the asset pricing theory's test Part I : on past and potential testability of the theory*. Journal of Financial Economics, **vol. 4**, n. 2, pages 129 à 179.

Richard Roll., Stephen Ross. (1980). *An empirical investigation of the Arbitrage pricing Theory* . The Journal of Finance, **vol. 35**, n. 5, pages 1073 à 1103.

Richard Roll., Stephen Ross. (1984). *The Arbitrage Pricing Theory approach to strategic portfolio planning*. The Journal of Finance, **vol. 40**, n. 3, pages 14 à 26.

Richard Roll., Stephen Ross. (1994). *On the cross-sectional relation between expected returns and betas*. The Journal of Finance, **vol. 49**, n. 1, pages 101 à 121.

Shlomo Yitzhaki., Edna Schechtman. (2013). *The Gini Methodology : A Primer on a Statistical Methodology*. New York : Springer.

Stephen Ross. (1976). *The arbitrage theory of capital asset pricing*. Journal of Economics Theory, **vol. 13**, n. 3, pages 341 à 360.

Téa Ouraga. (2019). *A note on Gini Principal Component Analysis*.Economics Bulletin, **vol. 39**, Issue 2.

William Sharpe. (1964). *Capital Asset prices : a theory of equilibrium under conditions of risk*. Journal of Finance, **vol. 19**, n. 3, pages 425 à 442.

CHAPITRE 5

GENERALIZED GINI LINEAR AND QUADRATIC DISCRIMINANT ANALYSIS

Sommaire

Résumé

Dans ce chapitre, une analyse discriminante linéaire (LDA) est effectuée au sens de Gini (GDA). La maximisation de la matrice brute de Gini généralisée entre les groupes permet de projeter les données sur les axes discriminants. Différentes méthodes sont étudiées, l'approche géométrique – basée sur une distance particulière – et l'approche probabiliste, qui consiste à utiliser la matrice généralisée de Gini intra-groupe afin de calculer la probabilité conditionnelle de classer les observations dans des groupes spécifiques. Des tests basés sur les statistiques U sont proposés afin de tester les variables discriminantes au lieu d'utiliser le test bien connu de Student qui requiert l'homoscédasticité. Ce chapitre a fait l'objet d'une publication, *C.f* Condevaux et al. (2020) dans lequel des simulations de Monte Carlo montrent la robustesse et la supériorité de la GDA sur les données contaminées par rapport à divers classificateurs linéaires tels que Logit, SVM et LDA.

Mots-clés : (R) Statistiques Multivariées ; Gini ; LDA ; Robustesse.

Abstract

In this chapter, a linear discriminant analysis (LDA) is performed in the Gini sense (GDA). Maximizing the generalized Gini gross between-group matrix allows the data to be projected onto discriminant axes. Different methods are investigated, the geometrical approach – based on a particular distance – and the probabilistic approach, which consists in employing the generalized Gini within-group matrix in order to compute the conditional probability of ranking observations in specific groups. Tests based on U-statistics are proposed in order to test for discriminant variables instead of using the well-known Student test that requires homoskedasticity. This chapter has been the subject of a paper, *i.e* Condevaux et al. (2020) in which Monte Carlo simulations show the robustness and the superiority of the GDA on contaminated data compared with various linear classifiers such as Logit, SVM and LDA.

Keywords : (R) Multivariate statistics ; Gini ; LDA ; Robustness.

5.1 Introduction

In 1912, Gini proposed the Gini Mean Difference (GMD) as a new measure to gauge inequality and disparity between individuals in a society. The GMD is based on the Manhattan distance (city-block distance), as a consequence, it provides a different way to measure inequality and variability compared with the usual variance based on the euclidean metrics.

Schechtman & Yitzhaki (1987) introduce the Gini covariance operator – coGini from now on – which can be seen as a mixture of the variance and Spearman's pure rank approaches. It is defined as the covariance between a random variable and its cumulative distribution function. The coGini operator has some appealing features. For instance in the case of regressions, Olkin and Yitzhaki (1992) and Yitzhaki & Schechtman (2013) point out that the ordinary least squares method can be abandoned for another technique in which the usual covariance operator is replaced by the coGini one. They propose a Gini regression, which has been shown to be robust to outliers. Indeed, the variance criterion may be misleading to handle a sample with extreme values or to deal with heavy-tailed distributions. Carcea & Serfling (2015) develop a Gini regression for time series analysis in which the usual auto-covariance function is replaced by the Gini auto-covariance function in order to deal with a new Box & Jenkins methodology robust to outliers. The Gini regression has also been extended to heteroskedastic data and fixed effects panel data, see respectively Charpentier *et al.* (2019a) and Ka & Mussard (2016).

In risk analysis, the use of the coGini operator has been proven to be useful, as shown for instance by Furman & Zitikis (2017). The coGini operator may be defined in such a way that it becomes close to Choquet integrals, which is very convenient to unify measures of inequality and measures of risk. This work is line with the use of rank-dependent measures of risks (and inequality) for decision-making under uncertainty, see Yaari (1987) and Yaari (1988).

In multidimensional statistics, the Gini index and the coGini operator are well-suited tools for classification purposes. Charpentier *et al.* (2019b) propose to employ the Gini correlation matrix in order to provide a principal component analysis robust to outliers. The space reduction is done by finding the eigenvector of the generalized Gini correlation matrix, this

technique being equivalent to the usual principal component analysis when the variables have Gaussian distributions. Also, Montanari (2004) and Calo (2006) propose to employ Gini's (multivariate) transvariation to provide a measure of group separability and to offer a linear discriminant analysis (LDA) in the Gini sense (GDA from now on). This GDA, compared with the usual Euclidean LDA, is also shown to be robust to outliers. This literature opens the way on LDA based on ℓ_1 norm. For instance, Li *et. al* (2015) propose to replace the Euclidean norm by the ℓ_1 norm in order to maximize the between-group variability in a given sample with the aid of an iterative algorithm (useful for image classification).

In this chapter, we start from the recognition that the variance criterion is not suitable for LDA because it captures a very restrictive notion of dispersion. It seems natural to analyze multidimensional data based on other norms, such as the ℓ_1 norm underlying the Gini index. We propose a matrix approach of the generalized Gini correlation indices in order to perform discriminant analyses, with two approaches, a geometrical one and a probabilistic one. Both techniques are appropriate to perform classifications when the data are drastically affected by outliers, **?**. we have the same conclusion in an application on two groups of efficient portfolios.

The chapter is organized as follows. Section 5.2 introduces the multidimensional Gini correlation and its decomposition. Section 5.3 exposes the linear and the quadratic discriminant analyses in the Gini sense. An application is performed in Section 5.4. Section 5.5 closes the paper.

5.2 Multidimensional Gini correlation

In this section, the well-known one-dimensional Gini correlation index is reviewed. Then, the Gini correlation matrix is introduced and its decomposition into within- and "gross between-group" Gini variability is proposed. [1] The maximization of the gross between-group variability is

1. In the Gini literature, the expression between-goup inequality (or between-group variability) is only used if the variables of the different groups do not overlap, see Yitzhaki & Schechtman (2013) for an overview of this literature. In the overlapping case, the gross between-group Gini term can be further decomposed into a stratification term (Yitzhaki & Lerman (1991), Yitzhaki (1994)) or an overlapping term (*i.e.* Gini's transvariation, see Dagum (1997)) plus a between-group inequality term (or between-group variability term) that accounts for the inequality (or variability) between the variables means.

maximized in order to derive the eigenvalue equation of the discriminant analysis in the Gini sense.

Gini correlation

Let $\mathbb{N}^*$ be the set of positive integers and $\mathbb{R}$ $[\mathbb{R}_{++}]$ the set of [positive] real numbers. Let $\mathcal{M}$ be the set of all $n \times K$ matrices $\boldsymbol{X} \equiv [x_{ik}]$ that describe n observations of a given sample on K dimensions such that $n \gg K$, $x_{ik} \in \mathbb{R}$ and $n, K \in \mathbb{N}^* \setminus \{1\}$, and $\mathbb{I}_n$ being the $n \times n$ identity matrix. Let $\mathbb{1}_n$ a n-dimensional vector of ones. The $n \times 1$ vectors representing each variable are expressed as $\mathbf{x}_{.k}$, for all $k \in \{1, \ldots, K\}$, such that $\mathbf{x}_{.k} \neq c\mathbb{1}_K$, with c a real constant. The $K \times 1$ vectors representing each individual i are expressed as $\mathbf{x}_{i.}$, for all $i \in \{1, \ldots, n\} \equiv \mathscr{S}$, where $\mathscr{S}$ is one given sample of the overall population. The arithmetic mean over each column (line) of the matrix $\boldsymbol{X}$ is given by $\bar{\mathbf{x}}_{.k}$ ($\bar{\mathbf{x}}_{i.}$).

The Gini mean difference parameter is defined as the expected absolute difference between two realizations of i.i.d. random variables :

$$\Delta := \mathbb{E}\{|X_1 - X_2|\}.$$

The one-dimensional Gini Mean Difference estimator, proposed by Schechtman & Yitzhaki (1987), is given by :

$$GMD(\mathbf{x}_{.k}) := \frac{4}{n}\sum_{i=1}^{n}(x_{ik} - \bar{\mathbf{x}}_{.k})(\hat{F}(x_{ik}) - \bar{F}_{\mathbf{x}_{.k}}),\ k = 1, \ldots, K,$$

where $\hat{F}$ is an estimator of the cumulative distribution function of $\mathbf{x}_{.k}$, and $\bar{F}_{\mathbf{x}_{.k}}$ its arithmetic mean. The one-dimensional GMD represents the variability of the variable $\mathbf{x}_{.k}$. The two-dimensional GMD measures the co-variability between two variables. The two-dimensional GMD between $\mathbf{x}_{.k}$ and $\mathbf{x}_{.\ell}$ is given by :

$$GMD(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}) := \frac{4}{n}\sum_{i=1}^{n}(x_{ik} - \bar{\mathbf{x}}_{.k})(\hat{F}(x_{i\ell}) - \bar{F}_{\mathbf{x}_{.\ell}}),\ k \neq \ell = 1, \ldots, K.$$

Alternatively, it is possible to employ the rank of observation i with respect to $\mathbf{x}_{.\ell}$ as an estimator of $nF(x_{i\ell})$:

$$\tilde{r}_{i\ell} := n\hat{F}(x_{i\ell}) = \begin{cases} \#\{x \leq x_{i\ell}\} & \text{if no ties} \\ \frac{\sum_{i=1}^{p}\#\{x \leq x_{i\ell}\}}{p} & \text{if } p \text{ ties } x_{i\ell}. \end{cases}$$

Then, a bias corrected estimator of the two-dimensional GMD is given by,

$$GMD(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}) = \frac{4}{n(n-1)} \sum_{i=1}^{n} (x_{ik} - \bar{\mathbf{x}}_{.k})(\tilde{r}_{i\ell} - \bar{\tilde{\mathbf{R}}}_{.\ell}),\ k \neq \ell = 1, \ldots, K,$$

with $\tilde{\mathbf{R}}_{.\ell}$ the rank vector of $\mathbf{x}_{.\ell}$ and $\bar{\tilde{\mathbf{R}}}_{.\ell}$ its arithmetic mean. In order to deal with a normalized coefficient of correlation, the Gini correlation coefficient (G-correlation frown now on) is studied. Since the GMD is not symmetric, two G-correlation indices are built according to their respective GMD :

$$GC(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) := \frac{GMD(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})}{GMD(\mathbf{x}_{.\ell})}\ ;\ GC(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}) := \frac{GMD(\mathbf{x}_{.k}, \mathbf{x}_{.\ell})}{GMD(\mathbf{x}_{.k})},$$

with $GC(\mathbf{x}_{.k}) = 1$ for all $k = 1, \ldots, K$. Following Yitzhaki (2003), the G-correlation is well-suited for the measurement of correlations in the case of non-normal distributions. One property of the G-correlation relies on exchangeable random variables. The random variables X and Y are said to be exchangeable if $F(X, Y) = F(Y, X)$. This implies exchangeability up to a linear transformation *i.e.* $F(X, Y) = F(aY + b, cX + d)$ such that $a, c > 0$, but the converse is not true, see Yitzhaki & Schechtman (2013). The main properties of the G-correlation index are as follows.

Property 5.2.1. – **Yitzhaki (2003) :** *For all* $k \neq \ell = 1, \ldots, K$ *:*
(i) $-1 \leq GC(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) \leq 1$.
(ii) *If the variables* $\mathbf{x}_{.\ell}$ *and* $\mathbf{x}_{.k}$ *are independent, then* $GC(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) = GC(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}) = 0$.
(iii) *For any given monotonic transformation* φ, $GC(\mathbf{x}_{.\ell}, \varphi(\mathbf{x}_{.k})) = GC(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})$ [*in the same manner as for Spearman's coefficient*].
(iv) *For any given linear transformation* φ, $GC(\varphi(\mathbf{x}_{.\ell}), \mathbf{x}_{.k}) = GC(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})$ [*in the same manner as for Pearson's coefficient*].
(v) *If* $\mathbf{x}_{.k}$ *and* $\mathbf{x}_{.\ell}$ *are exchangeable up to a linear transformation, then* $GC(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) = GC(\mathbf{x}_{.k}, \mathbf{x}_{.\ell})$.

Generalized Gini correlation matrix

Let $\mathbf{R}_{.k}$ be the decumulative rank vector of $\mathbf{x}_{.k}$, that is, the vector that assigns the smallest value (1) to the greater observation x_{ik} among all $i \in \{1, \ldots, n\}$ and so on. The sample covariance between $\mathbf{x}_{.k}$ and its rank

vector $\mathbf{x}_{.k}$, also called *the coGini operator* [2], is given by :

$$cov\Big(\mathbf{x}_{.k}, \mathbf{R}_{.k}\Big) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{\mathbf{x}}_{.k})(r_{ik} - \bar{\mathbf{R}}_{.k}).$$

On this basis, the one-dimensional generalized GMD is based on a single parameter ν such that :

$$GMD_{\nu}(\mathbf{x}_{.k}) := -\frac{2\nu}{n} cov\Big(\mathbf{x}_{.k}, \mathbf{R}_{.k}^{\nu-1}\Big),\ \nu > 1,$$

with $cov\Big(\mathbf{x}_{.\ell}, \mathbf{R}_{.k}^{\nu-1}\Big)$ being the generalized coGini operator. The two-dimensional generalized GMD is, for $k \neq \ell = 1, \ldots, K$,

$$GMD_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) := -\frac{2\nu}{n} cov\Big(\mathbf{x}_{.\ell}, \mathbf{R}_{.k}^{\nu-1}\Big),\ \nu > 1.$$

When $\nu = 2$, we retrieve the GMD as a special case, *i.e.* $GMD_2 = GMD$. Accordingly, two generalized Gini correlation coefficients arise :

$$GC_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) := \frac{GMD_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})}{GMD_{\nu}(\mathbf{x}_{.\ell})}\ ;\ GC_{\nu}(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}) := \frac{GMD_{\nu}(\mathbf{x}_{.k}, \mathbf{x}_{.\ell})}{GMD_{\nu}(\mathbf{x}_{.k})},$$

with $GC_{\nu}(\mathbf{x}_{.k}, \mathbf{x}_{.k}) = 1$ for all $k = 1, \ldots, K$. Following Yitzhaki & Schechtman (2013), the generalized GC_{ν}-correlation has the following properties.

Property 5.2.2. – **Yitzhaki & Schechtman (2013) :**
(i) $GC_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) \leq 1$.
(ii) *If the variables* $\mathbf{x}_{.\ell}$ *and* $\mathbf{x}_{.k}$ *are independent, for all* $k \neq \ell$, *then* $GC_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) = GC_{\nu}(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}) = 0$.
(iii) *For any given monotonic transformation* φ, $GC_{\nu}(\mathbf{x}_{.\ell}, \varphi(\mathbf{x}_{.k})) = GC_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})$.
(iv) *For any given linear transformation* φ, $GC_{\nu}(\varphi(\mathbf{x}_{.\ell}), \mathbf{x}_{.k}) = GC_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})$ [*in the same manner as for Pearson's coefficient*].
(v) *If* $\mathbf{x}_{.k}$ *and* $\mathbf{x}_{.\ell}$ *are exchangeable up to a linear transformation, then* $GC_{\nu}(\mathbf{x}_{.\ell}, \mathbf{x}_{.k}) = GC_{\nu}(\mathbf{x}_{.k}, \mathbf{x}_{.\ell})$.

Whenever $\nu \to 1$, the variability of the variables is mitigated so that GC_{ν} is close to zero (even if the variables exhibit a strong variability). On

2. Strictly speaking the coGini of two random variables X, Y is either given by $cov(X, G(Y))$ or $cov(Y, F(X))$, see Yitzhaki and Schechtman (2013), page 40.

the contrary, if $\nu \to \infty$ then GC_ν focuses on what happens at the lower tail of the distributions.

In this respect, the choice of ν enables one to mitigate or to improve the correlation GC_ν between the variables $\mathbf{x}_{.\ell}$ and $\mathbf{x}_{.k}$. Thereby, it is interesting to perform many Gini DA, with various values of ν in order to detect robustness in the significance of the potentially discriminant variables.

The starting point of this robustness check is to build the GC_ν-correlation matrix of $\boldsymbol{X}$.

Definition 5.2.1. – Gini correlation matrix : *The generalized Gini correlation matrix of $\boldsymbol{X}$ is a $K \times K$ matrix of the generalized Gini correlation between all couples $(\mathbf{x}_{.\ell}, \mathbf{x}_{.k})$ for all $k, \ell = 1, \ldots, K$:*

$$GC_\nu(\boldsymbol{X}) \equiv [GC_\nu(\mathbf{x}_{.k}, \mathbf{x}_{.\ell})]. \tag{5.1}$$

Following Property 5.2.1 (iv), it is possible to rescale the variables $\mathbf{x}_{.k}$ thanks to a linear transformation so that GC_ν remains constant. Therefore, let us standardized the matrix $\boldsymbol{X}$ into a $K \times K$ matrix $\boldsymbol{Z}_\nu$ as follows :

$$\boldsymbol{Z}_\nu \equiv \left[\frac{x_{ik} - \bar{\mathbf{x}}_{.k}}{GMD_\nu(\mathbf{x}_{.k})}\right]. \tag{5.2}$$

As a consequence, the variables $\mathbf{x}_{.k}$ are rescaled such that their generalized GMD is equal to unity. Now, we define the matrix of the centered decumulative rank vectors of $\mathbf{z}_{.k}$ (each $\mathbf{z}_{.k}$ being a column of $\boldsymbol{Z}_\nu$) :

$$\boldsymbol{R}^c_{\mathbf{z},\nu} \equiv [\mathbf{R}^{\nu-1}_{.1} - \overline{\mathbf{R}^{\nu-1}_{.1}}, \cdots, \mathbf{R}^{\nu-1}_{.K} - \overline{\mathbf{R}^{\nu-1}_{.K}}]. \tag{5.3}$$

The matrix $GC_\nu(\boldsymbol{X})$ may be expressed as the generalized Gini Mean Difference between all couples $(\mathbf{z}_{.k}, \mathbf{z}_{.\ell})$ for all $k, \ell = 1, \ldots, K$:

$$GC_\nu(\boldsymbol{X}) \equiv [GMD_\nu(\mathbf{z}_{.k}, \mathbf{z}_{.\ell})].$$

Indeed, let $\mathbf{z}^\intercal_{.k}$ be the transpose of $\mathbf{z}_{.k}$:

$$GMD_\nu(\mathbf{z}_{.k}, \mathbf{z}_{.\ell}) = -\frac{2\nu}{n(n-1)}\mathbf{z}^\intercal_{.k}[\mathbf{R}^{\nu-1}_{.\ell} - \overline{\mathbf{R}^{\nu-1}_{.\ell}}].$$

By Property 5.2.1, the rank vectors of each column of $\boldsymbol{X}$ are equal to those of $\boldsymbol{Z}_\nu$, then it comes that,

$$\begin{aligned} GMD_\nu(\mathbf{z}_{.k}, \mathbf{z}_{.\ell}) = -\frac{2\nu}{n(n-1)}\frac{[\mathbf{x}_{.k} - \bar{\mathbf{x}}_k]^\intercal[\mathbf{R}^{\nu-1}_{.\ell} - \overline{\mathbf{R}^{\nu-1}_{.\ell}}]}{GMD_\nu(\mathbf{x}_{.k})} &= \frac{GMD_\nu(\mathbf{x}_{.k}, \mathbf{x}_{.\ell})}{GMD_\nu(\mathbf{x}_{.k})} \\ &= GC_\nu(\mathbf{x}_{.k}, \mathbf{x}_{.\ell}). \end{aligned}$$

Finally, because $GMD_\nu(\mathbf{z}_{.k}) = 1$, denoting $\boldsymbol{Z}_\nu^\intercal$ the transpose of $\boldsymbol{Z}_\nu$, we have :

$$GC_\nu(\boldsymbol{X}) = [GC_\nu(\mathbf{x}_{.k}, \mathbf{x}_{.\ell})] = [GMD_\nu(\mathbf{z}_{.k}, \mathbf{z}_{.\ell})] = -\frac{2\nu}{n(n-1)}\boldsymbol{Z}_\nu^\intercal \boldsymbol{R}_{\mathbf{z},\nu}^c = GC_\nu(\boldsymbol{Z}_\nu).$$

Now, it is possible to investigate the subgroup decomposition of the generalized Gini correlation matrix $GC_\nu(\boldsymbol{X})$.

Decomposition of the generalized Gini correlation matrix

An important step before carrying out a discriminant analysis is to yield a decomposition of the multidimensional variability of $\boldsymbol{X}$ into within- and gross between-group variability. For that purpose, let the sample $\mathscr{S}$ be decomposed into G exclusive and exhaustive groups $\mathscr{S}_g$ of size n_g such that $g = 1, \ldots, G$ and $\mathscr{S}_g \cap \mathscr{S}_h = \emptyset$ for all $g \neq h = 1, \ldots, G$.

In this respect, the Gini variability between all couples $(\mathbf{z}_{.k}, \mathbf{z}_{.\ell})$ may be decomposed. We impose the following notations, on the one hand about the matrix $\boldsymbol{Z}_\nu$.

• Let $\mathbf{z}_{i.}^g$ be a $1 \times K$ vector, *i.e.* a line of $\boldsymbol{Z}_\nu$, representing the characteristics over K dimensions of an individual i in group $\mathscr{S}_g$.

• Let $\bar{\mathbf{z}} := (\bar{\mathbf{z}}_{.1}, \ldots, \bar{\mathbf{z}}_{.K}) = \mathbf{0}$ be the $1 \times K$ vector of the arithmetic mean of each variable k over the entire sample.

• Let $\bar{\mathbf{z}}^g$ be the $1 \times K$ vector of the mean of each variable k over group $\mathscr{S}_g$.

On the other hand, the notations about the ranks are the following.

• Let $\boldsymbol{R}_{\mathbf{z},\nu}^c \equiv [r_{c,ik}^{\nu-1}]$ being the matrix of the decumulative centered rank vectors of $\boldsymbol{Z}_\nu$ (powered by $\nu - 1$, see also Eq.(5.3)).

• Let $\bar{\mathbf{R}}^{\nu-1} = (\overline{\mathbf{R}_{.1}^{\nu-1}}, \ldots, \overline{\mathbf{R}_{.k}^{\nu-1}})$ be the $1 \times K$ vector of the arithmetic mean of each decumulative rank vector $\mathbf{R}_{.k}^{\nu-1}$ over sample $\mathscr{S}$.

• Let $[\mathbf{R}_{i.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}]$ be a $1 \times K$ vector which represents the ith line of matrix $\boldsymbol{R}_{\mathbf{z},\nu}^c$ (also denoted $[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}]$ if i is in group $\mathscr{S}_g$).

• Let $\bar{\mathbf{R}}_{g.}^{\nu-1}$ be the $1 \times K$ vector of the arithmetic mean of each decumulative rank vector $\mathbf{R}_{.k}^{\nu-1}$ over group $\mathscr{S}_g$.

In this respect it is possible to decompose the overall generalized Gini correlation matrix $GC_\nu(\boldsymbol{X})$ into within- and gross between-group correlation matrices.

Definition 5.2.2. – Gini correlation matrix within groups : *Let $\boldsymbol{R}_{\mathbf{z},w}^c$ be the $n \times K$ rank matrix in which each line i is given by $[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}_{g.}^{\nu-1}]$ and let $\boldsymbol{Z}_{\nu^w}^\intercal$ be*

the $n \times K$ matrix where each line i is given by $[\mathbf{z}_{i.}^{g} - \bar{\mathbf{z}}^{g}]$. The generalized Gini within-group correlation matrix is given by :

$$GC_{\nu,w}(\boldsymbol{X}) := -\frac{2\nu}{n(n-1)} \boldsymbol{Z}_{\nu^w}^{\intercal} \boldsymbol{R}_{\mathbf{z},w}^{c}. \tag{5.4}$$

Definition 5.2.3. – Gini correlation matrix between groups : *Let $\boldsymbol{R}_{\mathbf{z},b}^{c}$ be the $n \times K$ matrix where each line i is given by $[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}]$ for all $i \in \mathscr{S}_g$ and for all $g = 1, \ldots, G$, and let $\boldsymbol{Z}_{\nu^b}^{\intercal}$ be the $n \times K$ matrix where each line i of $\boldsymbol{Z}_{\nu}$ is replaced by $[\bar{\mathbf{z}}^{g} - \bar{\mathbf{z}}]$ for all $i \in \mathscr{S}_g$ and for all $g = 1, \ldots, G$, the generalized Gini gross-between-group correlation matrix is given by :*

$$GC_{\nu,b}(\boldsymbol{X}) := -\frac{2\nu}{n(n-1)} \boldsymbol{Z}_{\nu^b}^{\intercal} \boldsymbol{R}_{\mathbf{z},b}^{c}. \tag{5.5}$$

On this basis, a decomposition of the generalized Gini correlation matrix can be stated.

Proposition 5.2.1. *For a sample $\mathscr{S}$ decomposed into G exclusive and exhaustive groups $\mathscr{S}_g$ of size n_g such that $g = 1, \ldots, G$ and $\mathscr{S}_g \cap \mathscr{S}_h = \emptyset$ for all $g \neq h = 1, \ldots, G$. The generalized Gini correlation matrix $GC_{\nu}(\boldsymbol{X})$ is decomposable as follows :*

$$GC_{\nu}(\boldsymbol{X}) = GC_{\nu,w}(\boldsymbol{X}) + GC_{\nu,b}(\boldsymbol{X}).$$

If in addition $\bar{\boldsymbol{Z}}_{\nu}^{\,g}$ is the $n \times K$ matrix in which each line i corresponds to the mean $\bar{\mathbf{z}}^{g}$ of the ith observation of $\boldsymbol{Z}_{\nu}$ being in $\mathscr{S}_g$, then :

$$GC_{\nu,w}(\boldsymbol{X}) = (\boldsymbol{Z}_{\nu} - \bar{\boldsymbol{Z}}_{\nu}^{\,g})^{\intercal} \boldsymbol{R}_{\mathbf{z},\nu}^{c}$$
$$GC_{\nu,b}(\boldsymbol{X}) = \bar{\boldsymbol{Z}}_{\nu}^{\,g\intercal} \boldsymbol{R}_{\mathbf{z},\nu}^{c}.$$

Démonstration. See the Appendix. □

In contrast to the well-known ANOVA (variance analysis), the so-called ANOGI (Gini analysis) is very useful when dealing with outliers or contaminated data. Its robustness is a consequence of the fact that it uses the ℓ_1 norm. ANOGI is also of interest when the data deviate from the multivariate normal distribution, see Yitzhaki (2003). [3]

3. Proposition 5.2.1 outlines the fact that the generalized Gini gross between-group correlation matrix is obtained from $GC_{\nu,w}(\boldsymbol{X})$ as a residual. The Gini decomposition mimics the variance one : the within-group term measures the gap between each observation

Maximizing the gross between-group variability

Let us now investigate the use of the decomposition of the Gini correlation matrix $GC_\nu(\boldsymbol{X})$ in order to find a new subspace to perform a discriminant analysis.

Let $\mathbf{u}$ be a $K \times 1$ vector whose ℓ_1 norm is equal to unity $\|\mathbf{u}\|_1 := \sum_{k=1}^{K} |u_k| = 1$. The multidimensional Gini variability is given by the quadratic form $\mathbf{u}^\intercal GC_\nu(\boldsymbol{X})\mathbf{u}$. Thanks to the decomposition of the Gini correlation matrix $GC_\nu(\boldsymbol{X})$, the multidimensional Gini variability of sample $\mathscr{S}$ can be expressed as :

$$\mathbf{u}^\intercal GC_\nu(\boldsymbol{X})\mathbf{u} = \mathbf{u}^\intercal GC_{\nu,w}(\boldsymbol{X})\mathbf{u} + \mathbf{u}^\intercal GC_{\nu,b}(\boldsymbol{X})\mathbf{u}.$$

In this respect, the Gini discriminant analysis may be performed as the traditional DA based on the euclidean distance (variance).

The first way consists in maximizing the gross between-group variability as a share of the overall variability (i), whereas the second way consists in maximizing the gross between-group variability as a share of the within-group variability (ii).

Proposition 5.2.2.
(i) *The solution of* $\max \lambda := \frac{\mathbf{u}^\intercal GC_{\nu,b}(\boldsymbol{X})\mathbf{u}}{\mathbf{u}^\intercal GC_\nu(\boldsymbol{X})\mathbf{u}}$ *is given by the eigenvalue equation*

$$\left[GC_\nu(\boldsymbol{X}) + GC_\nu(\boldsymbol{X})^\intercal\right]^{-1}\left[GC_{\nu,b}(\boldsymbol{X}) + GC(\boldsymbol{X})_{\nu,b}^\intercal\right]\mathbf{u} = \lambda\mathbf{u}.$$

(ii) *The solution of* $\max \mu := \frac{\mathbf{u}^\intercal GC_{\nu,b}(\boldsymbol{X})\mathbf{u}}{\mathbf{u}^\intercal GC_{\nu,w}(\boldsymbol{X})\mathbf{u}}$ *is given by the eigenvalue equation*

$$\left[GC_{\nu,w}(\boldsymbol{X}) + GC_{\nu,w}(\boldsymbol{X})^\intercal\right]^{-1}\left[GC_{\nu,b}(\boldsymbol{X}) + GC(\boldsymbol{X})_{\nu,b}^\intercal\right]\mathbf{u} = \mu\mathbf{u}. \tag{5.6}$$

Démonstration. See the Appendix. □

Based on the two available eigenvalue equations, it is possible to project the data onto two different subspaces. Although these two approaches are equivalent, focus will be put on the within-group correlation matrix *i.e.* the so-called Mahalanobis metric.

in group $\mathscr{S}_g$ and the mean of group $\mathscr{S}_g$, while the gross between-group term measures the gap between the mean of each group $\mathscr{S}_g$ and the grand mean. Of course, because of the reranking term $\boldsymbol{R}^c_{\mathbf{z},\nu}$, we cannot see the overlap or stratification Gini term and the true between-group Gini term, see *e.g.* Yitzhaki & Lerman (1991) and Dagum (1997). Future researches could be done to show that the classification could be performed on the overlap term only as in Montanari (2004).

5.3 The generalized Gini Discriminant Analysis

In this section, we investigate the projection of the data onto the discriminant axes in order to predict the classification of the observations. The contributions of the individuals and the variables in the new subspace are computed. Then, some statistical tests are provided to assess the significance of the correlation between the variables and the discriminant axes.

Classification : the geometrical approach

In standard discriminant analysis, a first approach the so-called geometrical approach consists in computing the Euclidean distance between one individual and the gravity center of each group of the sample. The center of gravity $\bar{\mathbf{x}}^g$ of each group $g = 1, \ldots, G$ is a $1 \times K$ vector representing the arithmetic means of the variables computed over group $\mathscr{S}_g$, that is, $\bar{\mathbf{x}}^g = (\bar{\mathbf{x}}^g_{.1}, \ldots, \bar{\mathbf{x}}^g_{.K})$. The Euclidean distance of individual i from the center of gravity of group $\mathscr{S}_g$ is :

$$d^2(\mathbf{x}_{i.}, \bar{\mathbf{x}}^g) := (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)\mathbf{V}_w^{-1}(\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)^\intercal,$$

with $\mathbf{V}_w$ the $K \times K$ matrix of within-group covariances issued from the decomposition of the variance-covariance matrix of $\boldsymbol{X}$. The classification of individual i consists in minimizing the distance over all groups $\mathscr{S}_g$ and ranking i in the group $\mathscr{S}_{g^*}$ that minimizes the Euclidean distance.

Classification 5.3.1. – Euclidean distance : *The rule of classification based on the Euclidean distance is as follows :*

$$g_0^* = \underset{g=1,\ldots,G}{\arg\min}\, d^2(\mathbf{x}_{i.}, \bar{\mathbf{x}}^g).$$

The geometrical Gini approach is defined by mimicking the previous one. Instead of computing the Euclidean distance, the generalized Gini correlation is chosen instead,

$$d^1(\mathbf{x}_{i.}, \bar{\mathbf{x}}^g) := (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)GC_{\nu,w}(\boldsymbol{X})^{-1}(\mathbf{R}_{i.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1})^\intercal,$$

where $\mathbf{R}_{i.}^{\nu-1}$ is the ith line of matrix $\boldsymbol{R}^c_{\mathbf{z},\nu}$ *i.e.* the $1 \times K$ decumulative (powered) rank vector of observation i over each variable. [4]

4. Following Eq.(5.6) it is also possible to use $[GC_{\nu,w}(\boldsymbol{X}) + GC_{\nu,w}(\boldsymbol{X})^\intercal]^{-1}$ instead of $GC_{\nu,w}(\boldsymbol{X})^{-1}$.

Classification 5.3.2. – Geometric Gini 1 : *The rule of classification based on the generalized Gini correlation is as follows :*

$$g_1^* = \underset{g=1,\ldots,G}{\arg\min}\, d^1(\mathbf{x}_{i.}, \bar{\mathbf{x}}^g).$$

The previous rule of classification imitates the standard LDA approach based on the variance without rigorous motivation. Because discriminant analysis now relies on the GMD metric, it is possible to use the projector $\boldsymbol{B} \equiv [\mathbf{u}_{.1}, \cdots, \mathbf{u}_{.h}]$ where each $\mathbf{u}_{.k}$ represents the eigenvector stemming from the maximisation of the gross between-group Gini correlation matrix $GC_{\nu,b}(\boldsymbol{X})$ [Eq.(5.6)]. Indeed, projecting the observations such that $\boldsymbol{F} = \boldsymbol{Z}_\nu \boldsymbol{B}$, allows all GMDs to be computed in order to make a classification of the individuals in K groups. Let $\bar{\mathbf{f}}^g$ be the $1 \times K$ vector of the arithmetic means measured over each dimension k for group $\mathscr{S}_g$. In order to judge whether an individual belongs to group $\mathscr{S}_g$ we compute the projected data rescaled by the mean $\bar{\mathbf{f}}^g$ of group $\mathscr{S}_g$, that is, $\boldsymbol{F}_g \equiv [\mathbf{f}_{i.} - \bar{\mathbf{f}}^g]$. The rank matrix of $\boldsymbol{F}_g$ is equal to the rank matrix of $\boldsymbol{F}$ since the ranks remain invariant after rescaling the data. Let $\boldsymbol{R}_\mathbf{f} \equiv [r_{f,i\ell}^{\nu-1}]$ be the rank matrix of $\boldsymbol{F}$ (powered powered by $\nu - 1$).

Definition 5.3.1. *The two-dimensional generalized Gini Mean Difference* conditional *to group $\mathscr{S}_g$,*

$$GMD_{g,\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.\ell}) = -\frac{2\nu}{n_g(n_g-1)} \sum_{i=1}^{n} f_{ik} r_{f,i\ell}^{\nu-1},$$

provides the Gini co-variability between variables $\mathbf{f}_{.k}$ and $\mathbf{f}_{.\ell}$ when $\mathbf{f}_{.\ell}$ is rescaled such that all individuals are assumed to be in group $\mathscr{S}_g$.

An individual i in group $\mathscr{S}_g$ is correctly ranked in group $\mathscr{S}_g$, if for all $g = 1, \ldots, G$, his contribution to $GMD_{g,\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.\ell})$ given by

$$GMD_{ig,\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.\ell}) := -\frac{2\nu}{n(n-1)} f_{ik} r_{f,i\ell}^{\nu-1}$$

is the lowest possible.

Classification 5.3.3. – Geometric Gini 2 : *The rule of classification based on the conditional rank matrix is as follows :*

$$g_2^* = \underset{g=1,\ldots,G}{\arg\min}\, GMD_{ig,\nu}(\mathbf{f}_{.k}, \mathbf{f}_{.\ell}).$$

Another rule of classification consists in measuring the ℓ_1 distance between each individual position $\mathbf{f}_{i.}$ in the new subspace and the mean of each group $\bar{\mathbf{f}}^g$.

Classification 5.3.4. – Geometric Gini 3 : *The rule of classification based on the ℓ_1 distance is as follows :*

$$g_3^* = \underset{g=1,\ldots,G}{\arg\min} \|\mathbf{f}_{i.}, \bar{\mathbf{f}}^g\|_1.$$

Classification : the Gaussian approach

The usual LDA is performed by using the Gaussian assumption. The probability for an individual to be ranked in group $\mathscr{S}_g$ is :

$$\mathbb{P}[i \in \mathscr{S}_g] = \frac{1}{|\mathbf{V}_{w,g}|^{1/2} (2\pi)^{1/2}} \exp\left\{-\frac{1}{2} (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g) \mathbf{V}_{w,g}^{-1} (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)^\intercal\right\}, \quad (5.7)$$

where $\mathbf{V}_{w,g}$ is the $K \times K$ variance-covariance matrix of $\boldsymbol{X}$ computed over group $\mathscr{S}_g$. In this respect, a confusion matrix [5] may be derived in order to assess the quality of the ranking displayed by the discriminant axis $\mathbf{f}_{.k}$. It is also possible to deduce from Eq.(5.7) the usual LDA based either on homoskedasticity or heteroskedasticity.

Classification 5.3.5. – Homoskedastic LDA : *The homoskedastic LDA rule of classification is :*

$$g_4^* = \underset{g=1,\ldots,G}{\arg\min} \left\{-2\ln(n_g/n) + (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g) \hat{\mathbf{V}}_w^{-1} (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)^\intercal\right\},$$

with $\hat{\mathbf{V}}_w$ an unbiased estimator of $\mathbf{V}_w$,

$$\hat{\mathbf{V}}_w = \frac{1}{n-G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)^\intercal (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g).$$

Classification 5.3.6. – Heteroskedastic LDA : *The heteroskedastic LDA rule of classification is :*

$$g_5^* = \underset{g=1,\ldots,G}{\arg\min} \left\{-2\ln(n_g/n) + \ln(|\mathbf{V}_{w,g}|) + (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g) \mathbf{V}_{w,g}^{-1} (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)^\intercal\right\}.$$

5. A confusion matrix allows one to compute the performance of a classifier. It consists in measuring the rate of good (bad) classification and also the rate of false negative and false positive.

As studied by Baccini *et al.* (1996) in the case of principal component analysis, it is convenient to replace the standard deviation by a GMD estimator when outliers occur in the sample. On this basis, we derive two rules of classification based on the Gini variability within groups. Let $\widehat{GMD}_{w,\nu}(\boldsymbol{X})$ be the within-group variability defined as the weighted mean of the generalized GMD within groups :

$$\widehat{GMD}_{w,\nu}(\boldsymbol{X}) = \frac{n_g}{n-G}\sum_{g=1}^{G} GMD_{\nu,g}(\boldsymbol{X}).$$

Classification 5.3.7. – Homoskedastic Gini : *The homoskedastic GDA rule of classification is :*

$$g_6^* = \underset{g=1,\ldots,G}{\arg\min}\left\{ -2\ln(n_g/n) + (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)\,\widehat{GMD}_{w,\nu}(\boldsymbol{X})^{-1}\,(\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)^{\intercal} \right\}.$$

Classification 5.3.8. – Heteroskedastic Gini : *The heteroskedastic GDA rule of classification is :*

$$g_7^* = \underset{g=1,\ldots,G}{\arg\min}\left\{-2\ln(n_g/n) + \ln(|GMD_{\nu,g}(\boldsymbol{X}^g)|) + (\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)\,GMD_{\nu,g}^{-1}\,(\mathbf{x}_{i.} - \bar{\mathbf{x}}^g)^{\intercal}\right\},$$

with $GMD_{\nu,g}(\boldsymbol{X}^g)$ the $K \times K$ matrix of the generalized Gini Mean Difference computed over group $\mathscr{S}_g$.

With Gaussian random variables, it is possible to show that the LDA is a special case of the classification based on the Gini Mean Difference.

Proposition 5.3.1. *The two following assertions hold.*
(i) *If $\boldsymbol{X}^g = \boldsymbol{Z}_\nu^g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_g^2)$ for all $g = 1, \ldots, G$, then the homoskedastic LDA rule is equivalent to the homoskedastic GDA rule for all $i = 1, \ldots, n_g$:*

$$g_4^* = g_6^*.$$

(ii) *If $\boldsymbol{X}^g = \boldsymbol{Z}_\nu^g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_g^2)$ such that $\boldsymbol{\Sigma}_g^2 \neq \boldsymbol{\Sigma}_h^2$ for some $g \neq h \in \{1, \ldots, G\}$, then the heteroskedastic LDA rule is equivalent to the heteroskedastic GDA rule for all $i = 1, \ldots, n_g$:*

$$g_5^* = g_7^*.$$

Démonstration. Since by definition $\boldsymbol{Z}_{\nu}^{g}$ is standardized, then $GC_{\nu}(\boldsymbol{Z}_{\nu}^{g}) = GMD_{\nu}(\boldsymbol{Z}_{\nu}^{g})$ and $\mathbf{V}_{w} = \boldsymbol{\rho}_{w}$ where $\boldsymbol{\rho}_{w}$ is Pearson correlation coefficient matrix within groups. From Yitzhaki & Schechtman (2013) (Chapter 6), whenever $\boldsymbol{X} = (X_1, \ldots, X_K)$ have some K-variate Gaussian distributions then the generalized Gini correlation coefficient $GC_{\nu}(X_i, X_j) = \rho(X_i, X_j)$ for all $i, j = 1, \ldots, K$, for all $\nu > 1$. Then, we deduce in case (i) that $\widehat{GMD}_{w,\nu}(\boldsymbol{Z}_{\nu}) = \hat{\mathbf{V}}_{w}$ and so $g_4^* = g_6^*$. In case (ii), $\mathbf{V}_{w,g} = GMD_{\nu,g}(\boldsymbol{Z}_{\nu}^{g})$ for all $g = 1, \ldots, G$ and so $g_5^* = g_7^*$, which ends the proof.

□

Contribution of the individuals and the variables

Let the basis $\mathscr{B} := \{\mathbf{u}_{.1}, \ldots, \mathbf{u}_{.h}\}$ such that $h \leq K$ and $\boldsymbol{B}$ the $K \times h$ matrix such that $\boldsymbol{B} := (\mathbf{u}_{.1}, \ldots, \mathbf{u}_{.h})$, with $\mathbf{u}_{.1}$ the eigenvector corresponding to the highest eigenvalue μ_1, and so on. The coordinates of the individuals in the new subspace $\{\mathbf{f}_{.1}, \ldots, \mathbf{f}_{.h}\}$ issued from the projector $\boldsymbol{B}$ are :

$$\boldsymbol{F} = \boldsymbol{Z}_{\nu}\boldsymbol{B}.$$

The $n \times h$ matrix $\boldsymbol{F}$ allows the individuals to be analyzed on different discriminant axes $\mathbf{f}_{.1}, \ldots, \mathbf{f}_{.h}$. In order to find the individuals that contribute the most to each discriminant axis, their contribution to the Gini variability over each discriminant axis must be computed.

In the standard linear discriminant analysis, the quality of the discriminant axis may be examined by computing the distance between the gravity center (mean) of two groups $\mathscr{S}_g$ and $\mathscr{S}_h$ on each discriminant axis $\mathbf{f}_{.k}$. Let $\bar{\mathbf{f}}_{.k}^{g}$ and $\bar{\mathbf{f}}_{.k}^{h}$ the arithmetic means of $\mathbf{f}_{.k}$ measured on groups $\mathscr{S}_g$ and $\mathscr{S}_h$ respectively, with s_g^2 and s_h^2 their empirical variance. Let f_{ik} be the coordinate of individual i on axis $\mathbf{f}_{.k}$. Assuming that $f_{ik} \sim \mathcal{N}$ for all $i = 1, \ldots, n$ and that the sampling error is α, then the significance of the distance between groups $\mathscr{S}_g$ and $\mathscr{S}_h$ on the discriminant axis $\mathbf{f}_{.k}$ is given by :

$$T_{\mathbf{f}_{.k}} := \frac{\bar{\mathbf{f}}_{.k}^{g} - \bar{\mathbf{f}}_{.k}^{h}}{\sqrt{\frac{1}{n_g} + \frac{1}{n_h}} \cdot \sqrt{\frac{n_g s_g^2 + n_h s_h^2}{n_g + n_h - 2}}} \sim \mathcal{T}(n_g + n_h - 2),$$

where $\mathcal{T}$ is the Student distribution. If $|T_{\mathbf{f}_{.k}}| \geq \mathcal{T}_{\alpha/2}$, then the ℓ_1 distance $|\bar{\mathbf{f}}_{.k}^{g} - \bar{\mathbf{f}}_{.k}^{h}|$ between the mean groups $\mathscr{S}_g$ and $\mathscr{S}_h$ on the discriminant axis

$\mathbf{f}_{.k}$ is significant *i.e.* the mean gap is different from 0 (the percentile $\mathcal{T}_{\alpha/2}$ is given at the $\alpha/2$ level). Of course, the reliability of this test depends on the hypothesis of homoskedasticity. Thereby, it cannot systematically provide the discriminant variables $\mathbf{x}_{.k}$ associated with each axis $\mathbf{f}_{.k}$.

For this purpose, let us propose other statistics based on the Gini metric in order to measure the significance of the observations and of the variables on each axis $\mathbf{f}_{.k}$. The first one is the contribution of one observation to the variability of one given discriminant axis.

Definition 5.3.2. – Absolute contribution of the observations : *The absolute contribution of observation i to the generalized Gini variability of the discriminant axis $\mathbf{f}_{.k}$ is :*

$$ACT_{ik} := -\frac{2\nu}{n(n-1)}\frac{f_{ik}r_{cf,ik}^{\nu-1}}{GMD_\nu(\mathbf{f}_{.k})},$$

with $r_{cf,ik}^{\nu-1}$ the (powered) centered decumulative rank of observation i on axis $\mathbf{f}_{.k}$.

The absolute contribution of each i to the generalized Gini mean difference of $\mathbf{f}_{.k}$ is such that $ACT_{ik} \in [0,1]$ and $\sum_{i=1}^{n} ACT_{ik} = 1$. It provides the contribution of each observation i to $GMD_\nu(\mathbf{f}_{.k})$.

Let us now define another basis of $\mathbb{R}^n$ in order to further analyze the variables in a new subspace. Let us define $\{\mathbf{a}_{.1}, \ldots, \mathbf{a}_{.h}\}$ with $h \leq K$ the set resulting from the vectors $\mathbf{f}_{.k}$ such that :

$$\mathbf{a}_{.k} = \frac{\mathbf{f}_{.k}}{GMD_\nu(\mathbf{f}_{.k})}.$$

The set $\{\mathbf{a}_{.1}, \ldots, \mathbf{a}_{.h}\}$ is an orthonormal basis since $\mathbf{a}_{.k} \perp \mathbf{a}_{.\ell}$ for all $k \neq \ell \in \{1, \ldots, h\}$ with $GMD_\nu(\mathbf{a}_{.k}) = 1$ for all $k \in \{1, \ldots, h\}$. Let $\mathbf{A} := [\mathbf{a}_{.1}, \ldots, \mathbf{a}_{.h}]$ be the $n \times h$ matrix, which may serve to project the variables $\mathbf{z}_{.k}$ onto $\mathbb{R}^n$. If $\boldsymbol{V}$ denotes the $K \times h$ matrix of the coordinates of the variables in the new subspace, then

$$\boldsymbol{V} := -\frac{2\nu}{n(n-1)}\mathbf{A}^\intercal \boldsymbol{R}^c_{\mathbf{z},\nu}.$$

Instead of using the Pearson correlation coefficient between $\mathbf{z}_{.k}$ and $\mathbf{a}_{.\ell}$ to assess the contribution of the variables $\mathbf{z}_{.k}$ to each axis $\mathbf{a}_{.\ell}$, it is possible to use the Gini variability. Indeed, it is noteworthy that,

$$\boldsymbol{V} \equiv [GMD_\nu(\mathbf{a}_{.\ell}, \mathbf{z}_{.k})].$$

As a consequence, it is possible to determine whether or not a variable $\mathbf{z}_{.k}$ is discriminant for a given axis $\mathbf{a}_{.\ell}$.

Definition 5.3.3. – Absolute contribution of the variables : *The absolute contribution of variable $\mathbf{z}_{.k}$ to the generalized Gini variability of the discriminant axis $\mathbf{a}_{.\ell}$ is :*

$$ACT_{\ell k} = GMD_{\nu}(\mathbf{a}_{.\ell}, \mathbf{z}_{.k}).$$

If the variability between $\mathbf{a}_{.\ell}$ and $\mathbf{z}_{.k}$ is strong, *i.e.* $GMD_{\nu}(\mathbf{a}_{.\ell}, \mathbf{z}_{.k})$ exhibits important values, then the variable $\mathbf{z}_{.k}$ is said to be discriminant. In order to get a more intuitive interpretation of $ACT_{\ell k}$, we can equivalently use the generalized Gini correlation index.

Proposition 5.3.2. *For all $\ell, k \in \{1, \ldots, K\}$ the absolute contribution of variable $\mathbf{z}_{.k}$ to the generalized Gini variability of the discriminant axis $\mathbf{f}_{.\ell}$ is given by :*

$$ACT_{\ell k} = GC_{\nu}(\mathbf{f}_{.\ell}, \mathbf{z}_{.k}).$$

Démonstration. See the Appendix. □

Fallowing Proposition 5.3.2, a variable $\mathbf{z}_{.k}$ is said to be discriminant if its generalized Gini correlation is important with respect to axis $\mathbf{f}_{.\ell}$. To gauge to which extent this correlation is strong a U-statistics test is performed. Setting $\boldsymbol{V} \equiv [v_{\ell k}]$, it is possible to test directly for the significance of the elements $v_{\ell k}$ in order to capture the most discriminant variables $\mathbf{z}_{.k}$ on $\mathbf{f}_{.\ell}$.

Proposition 5.3.3. *Let $\widetilde{U}_{\ell k} := GMD_{\nu}(\mathbf{f}_{.\ell}, \mathbf{z}_{.k})$ with $U^0_{\ell k} := \mathbb{E}[U_{\ell k}]$ and $U_{\ell} := GMD_{\nu}(\mathbf{f}_{.k})$, then following assertions hold :*
(i) *The elements $v_{\ell k}$ of $\boldsymbol{V}$ are U-statistics :*

$$U_{\ell k} := \frac{\widetilde{U}_{\ell k}}{U_{\ell}} = v_{\ell k}.$$

(ii) *For $n \to \infty$, $\sqrt{n}(U_{\ell k} - U^0_{\ell k}) \overset{a}{\sim} \mathcal{N}$.*

Démonstration. From Yitzhaki & Schechtman (2013), $U_{\ell k}$ is an unbiased and consistent estimator of $U^0_{\ell k}$. From Theorem 10.4 in Yitzhaki & Schechtman (2013) (Chapter 10), the result follows. □

Therefore, from Proposition 5.3.3, it is possible to test for :

$$\left\| \begin{array}{l} H_0 : U^0_{\ell k} = 0 \\ H_1 : U^0_{\ell k} \neq 0. \end{array} \right.$$

Let $\hat{\sigma}_{\ell k}$ be the Jackknife estimator of the standard deviation of $U_{\ell k}$, then the null hypothesis may be tested making use of the following statistics :

$$\frac{U_{\ell k}}{\hat{\sigma}_{\ell k}} \overset{a}{\sim} \mathcal{N}(0,1).$$

5.4 Application

We propose an application of discriminant analysis in portfolio management. Based on the $Equitics^{®}$ methodology developed by Vigeo Eiris, a responsible governance score is used to rank CAC-40 companies. This score is given by the $CAC40^{®}$ Governance index of Euronext.

On the basis of this index, two groups of efficient portfolios are formed. The first group, called green portfolios, is derived from assets (8) with a good score for responsible governance. A portfolio in this group has a value of 1. The second group, called non-green portfolios, is derived from assets (8) with a poor responsible governance score. All portfolios in this group of assets take the value 0.

The selected green assets are : AIR LIQUIDE, CREDIT AGRICOLE, LEGRAND, PUBLICIS, SODEXO, STMICRO, VALEO & VEOLIA and the selected non-green assets are : AIRBUS, ARCELOR MITTAL, BNP-PARIBAS, DASSAULT, LVMH, SAFRAN, SOCIETE GENERALE & TOTAL. The efficient frontiers are shown below :

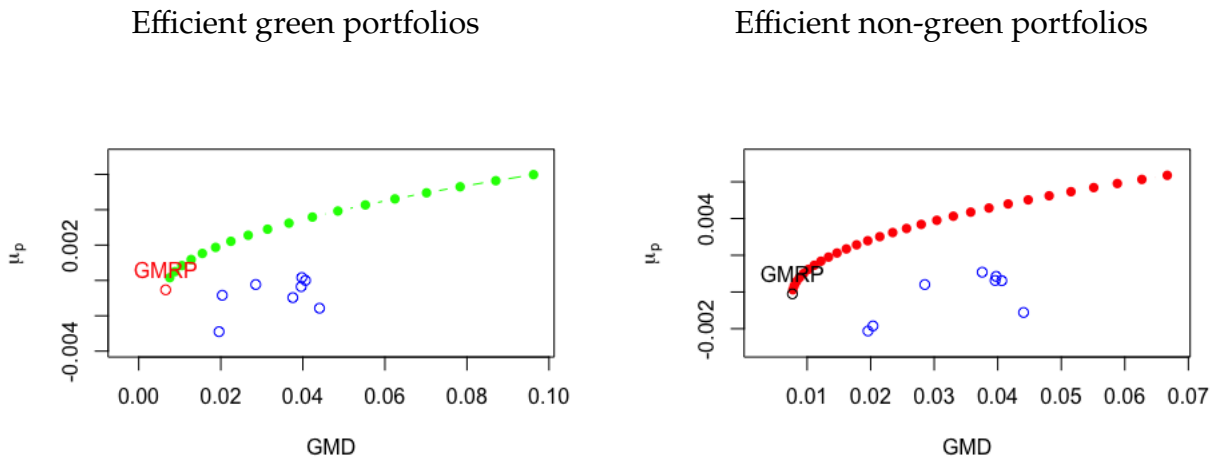


FIGURE 5.1 – Efficient frontiers

The expression $GMRP$ on Figure 5.1 represents Global Minimal Risk [6] Portfolio. The first eighteen portfolios with a positive expected return are selected from each group.

6. The risk is measured by the GMD.

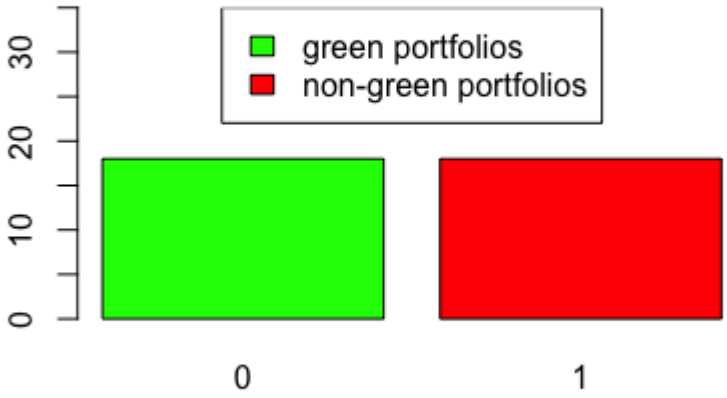


FIGURE 5.2 – Repartition by group

We have eighteen green portfolios and eighteen non-green portfolios represented on Figure 5.2.

These portfolios are then described by five performance ratios : sharpe ratio, Treynor ratio, Information ratio (IR), Black-Treynor ratio and Sortino ratio.

Portfolios	Sharpe	Treynor	IR	Black-Treynor	Sortino
1	0.12466	0.00181	0.05728	0.00054	0.07321
1	0.14356	0.00261	0.08410	0.00122	0.11904
1	0.15305	0.00356	0.09986	0.00201	0.15179
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
0	0.11343	0.00133	0.06789	0.00037	0.05655
0	0.14102	0.00178	0.11520	0.00070	0.10258
0	0.16470	0.00222	0.15068	0.00103	0.15110

TABLE 5.1 – Head of data by group

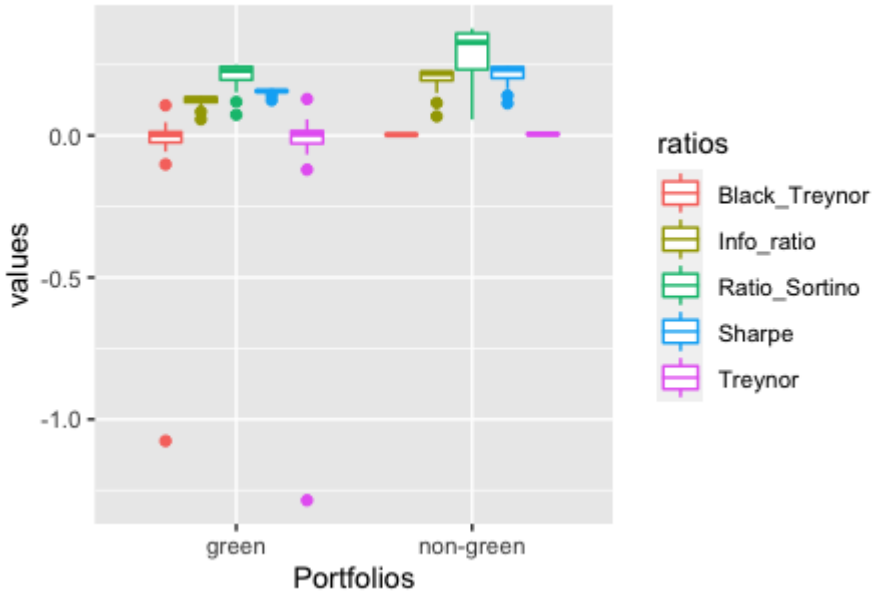


FIGURE 5.3 – Boxplot by group of portfolios

We suspect the presence of outliers in the set of ratios for the green portfolios, particularly for the Treynor and Black-Treynor ratios. With regard

to non-green portfolios, there is a suspicion of extreme observations at the level of Information and Sharpe ratios. In addition, non-green portfolios seem to outperform green portfolios according to Sortino, Sharpe and information criteria.

The dataset is interesting since there are highly correlated ratios (variables) and ratios with low correlations as can be seen in the Pearson correlation matrix given in Table 5.2 :

	Sharpe	**Treynor**	**IR**	**Black-Treynor**	**Sortino**
Sharpe	**1.000**	0.127	0.975	0.126	0.894
Treynor	0.127	**1.000**	0.085	0.99	0.010
IR	0.975	0.085	**1.000**	0.084	0.904
Black-Treynor	0.126	0.99	0.084	**1.000**	0.009
Sortino	0.894	0.010	0.904	0.009	**1.000**

TABLE 5.2 – Correlation matrix

We perform a comparison between LDA and GDA algorithms on the dataset Table 5.1. We compare the methods according to the overall prediction error rate of each algorithm. The algorithm with the lowest overall prediction error rate has been obtained by choosing the optimal value of ν which minimizes the error rate, that is, $\nu = 2.5$. The results of this application are mentioned in the table below.

In the geometrical approaches, the Gini approaches g_1 and g_3 with a parameter $\nu = 2.5$ have the same overall error rate as the conventional approach g_0, i.e 2%.

Method	Group 1	Group 2	Overall error rate
Euclidean distance	0%	5%	2%
Geometric Gini 1 (g_1) & $\nu = 2.5$	11%	88%	50%
Geometric Gini 2 & $\nu = 2.5$	5%	0%	2%
Geometric Gini 3 & $\nu = 2.5$	0%	5%	2%

TABLE 5.3 – Error rate : Classification by geometrical approach

In the probabilistic cases (Gaussian approaches), we can distinguish homoskedastic (g_4 and g_6) and heteroskedastic rules of classification (g_5 and g_7). The Gini approaches g_6 and g_7 overperform the standard approaches g_6 and g_7 with a parameter $\nu = 1.5$.

Method	Group 1	Group 2	Overall error rate
Homoskedastic LDA	$0\ \%$	100%	50%
Homoskedastic GDA & $\nu = 1.5$	88%	5%	47%
Heteroskedastic LDA	11%	0%	5%
Heteroskedastic GDA & $\nu = 1.5$	0%	5%	2%

TABLE 5.4 – Error rate : Classification by gaussian approach

Discussion

We wish to observe the performance of the algorithms if extreme portfolios existed in this dataset. For example, how investment funds that would

have performed during a period of crisis. To do so, we randomly contaminate two portfolios in each portfolio group, which represents 11% of the dataset. The results are as follows :

Method	Group 1	Group 2	Overall error rate
Euclidean distance	5%	22%	13%
Geometric Gini 1 & $\nu = 5$	100%	0%	50%
Geometric Gini 2 & $\nu = 5$	38%	5%	22%
Geometric Gini 3 & $\nu = 5$	0%	11%	5%

TABLE 5.5 – Error rate : Classification by geometrical approach with contamination

In this approach, the geometric Gini 3 (g_3) with a parameter $\nu = 5$ has the best overall error rate and a gap of 8% with the classical method (g_0). Nevertheless, the performance of all algorithms using this approach has decreased.

Method	Group 1	Group 2	Overall error rate
Homoskedastic LDA	$88\ \%$	$0\ \%$	44%
Homoskedastic GDA & $\nu = 5$	0%	100%	50%
Heteroskedastic LDA	0.05%	22%	13%
Heteroskedastic GDA & $\nu = 5$	0%	0%	0%

TABLE 5.6 – Error rate : Classification by gaussian approach with contamination

The Heteroskedastic GDA & $\nu = 5$ perform better than other probabilistic models in the presence of extreme portfolios in the dataset with an overall error rate 0%. It should also be noted that both Gini models have an error rate equal to 0% in group 1.

5.5 Conclusion

In this chapter, we have shown that the Gini correlation matrix is decomposable into within- and gross between-group variability. Accordingly,

maximizing the generalized gross between-group Gini correlation matrix allows one to project the data onto discriminant axes. Among the geometrical approaches, three rules of classification based on the generalized Gini index have been proposed. In the application, the rules g_2 seems to be efficient with a same overall error rate of classification that g_0, g_3 and g_7 in the absence of extreme portfolios in the dataset. However, as shown in the Application, the gaussian classification rule g_7 (Heteroskedastic GDA) performs better when rows (portfolios) are affected by outliers with zero classification error.

Appendix : Proofs

Proof of Proposition 5.2.1 :
The Gini variability decomposition is :

$$\begin{aligned}
GC_\nu(\boldsymbol{X}) &= -\frac{2\nu}{n(n-1)}\boldsymbol{Z}_\nu^\intercal \boldsymbol{R}_{\mathbf{z},\nu}^c \\
&= -\frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\mathbf{z}_{i.}^g - \bar{\mathbf{z}})^\intercal[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}] \\
&= -\frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\mathbf{z}_{i.}^g + \bar{\mathbf{z}}^g - \bar{\mathbf{z}}^g - \bar{\mathbf{z}})^\intercal[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}] \\
&= -\frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\mathbf{z}_{i.}^g - \bar{\mathbf{z}}^g)^\intercal[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}] \\
&\quad - \frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\bar{\mathbf{z}}^g - \bar{\mathbf{z}})^\intercal[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}].
\end{aligned}$$

Since,

$$\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\mathbf{z}_{i.}^g - \bar{\mathbf{z}}^g)^\intercal[\bar{\mathbf{R}}^{\nu-1}] = \sum_{i=1}^{n_g}\sum_{g=1}^{G}(\mathbf{z}_{i.}^g - \bar{\mathbf{z}}^g)^\intercal[\bar{\mathbf{R}}_{g.}^{\nu-1}] = \mathbf{0} \tag{5.8}$$

then,

$$\begin{aligned}
GC_\nu(\boldsymbol{X}) &= -\frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\mathbf{z}_{i.}^g - \bar{\mathbf{z}}^g)^\intercal[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}_{g.}^{\nu-1}] \\
&\quad - \frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\bar{\mathbf{z}}^g - \bar{\mathbf{z}})^\intercal[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}^{\nu-1}] \\
&= -\frac{2\nu}{n(n-1)}\boldsymbol{Z}_{\nu^w}^\intercal \boldsymbol{R}_{\mathbf{z},w}^c - \frac{2\nu}{n(n-1)}\boldsymbol{Z}_{\nu^b}^\intercal \boldsymbol{R}_{\mathbf{z},b}^c \\
&= GC_{\nu,w}(\boldsymbol{X}) + GC_{\nu,w}(\boldsymbol{X}).
\end{aligned}$$

Let $\bar{\boldsymbol{Z}}_\nu{}^g$ is the $n \times K$ matrix in which each line i corresponds to the mean $\bar{\mathbf{z}}^g$ of the ith observation of $\boldsymbol{Z}_\nu$ being in $\mathscr{S}_g$, then by Eq.(5.8) :

$$\begin{aligned}
-\frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\mathbf{z}_{i.}^g - \bar{\mathbf{z}}^g)^\intercal[\mathbf{R}_{ig.}^{\nu-1} - \bar{\mathbf{R}}_{g.}^{\nu-1}] &= (\boldsymbol{Z}_\nu - \bar{\boldsymbol{Z}}_\nu{}^g)^\intercal \boldsymbol{R}_{\mathbf{z},\nu}^c \\
&= GC_{\nu,w}(\boldsymbol{X}).
\end{aligned}$$

By Eq.(5.8) again :

$$-\frac{2\nu}{n(n-1)}\sum_{i=1}^{n_g}\sum_{g=1}^{G}(\bar{\mathbf{z}}^g-\bar{\mathbf{z}})^\intercal[\mathbf{R}_{ig.}^{\nu-1}-\bar{\mathbf{R}}^{\nu-1}]=\bar{\boldsymbol{Z}}_\nu^{g\intercal}\boldsymbol{R}_{\mathbf{z},\nu}^c$$
$$=GC_{\nu,b}(\boldsymbol{X}),$$

and this ends the proof.

Proof of Proposition 5.2.2 :
(i) Since $GC_{\nu,w}(\boldsymbol{X})$ and $GC_{\nu,b}(\boldsymbol{X})$ are not symmetric matrices, it comes that maximizing

$$\lambda:=\frac{\mathbf{u}^\intercal GC_{\nu,b}(\boldsymbol{X})\mathbf{u}}{\mathbf{u}^\intercal GC_{\nu}(\boldsymbol{X})\mathbf{u}}$$

yields,

$$\frac{[GC_{\nu,b}(\boldsymbol{X})+GC_{\nu,b}(\boldsymbol{X})^\intercal]\mathbf{u}\,[\mathbf{u}^\intercal GC_{\nu}(\boldsymbol{X})\mathbf{u}]-[\mathbf{u}^\intercal GC_{\nu,b}(\boldsymbol{X})\mathbf{u}]\,[GC_{\nu}(\boldsymbol{X})+GC_{\nu}(\boldsymbol{X})^\intercal]\mathbf{u}}{[\mathbf{u}^\intercal GC_{\nu}(\boldsymbol{X})\mathbf{u}]^2}=0.$$

This entails the following eigenvalue equation :

$$[GC_{\nu}(\boldsymbol{X})+GC_{\nu}(\boldsymbol{X})^\intercal]^{-1}\left[GC_{\nu,b}(\boldsymbol{X})+GC(\boldsymbol{X})_{\nu,b}^\intercal\right]\mathbf{u}=\lambda\mathbf{u}.$$

(ii) The other option is to maximize the gross between-group variability as a share of the within-group one, *i.e.*

$$\mu:=\frac{\mathbf{u}^\intercal GC_{\nu,b}(\boldsymbol{X})\mathbf{u}}{\mathbf{u}^\intercal GC_{\nu,w}(\boldsymbol{X})\mathbf{u}},$$

then we get :

$$\frac{[GC_{\nu,b}(\boldsymbol{X})+GC_{\nu,b}(\boldsymbol{X})^\intercal]\mathbf{u}\,[\mathbf{u}^\intercal GC_{\nu,w}(\boldsymbol{X})\mathbf{u}]-[\mathbf{u}^\intercal GC_{\nu,b}(\boldsymbol{X})\mathbf{u}]\,[GC_{\nu,w}(\boldsymbol{X})+GC_{\nu,w}(\boldsymbol{X})^\intercal]\mathbf{u}}{[\mathbf{u}^\intercal GC_{\nu,w}(\boldsymbol{X})\mathbf{u}]^2}=0.$$

Therefore, a Mahalanobis metric in the Gini sense is obtained thanks to the following eigenvalue equation,

$$[GC_{\nu,w}(\boldsymbol{X})+GC_{\nu,w}(\boldsymbol{X})^\intercal]^{-1}\left[GC_{\nu,b}(\boldsymbol{X})+GC(\boldsymbol{X})_{\nu,b}^\intercal\right]\mathbf{u}=\mu\mathbf{u}.$$

Proof of Proposition 5.3.2 :
From (5.8), since $\bar{\mathbf{a}}_{.\ell} = \bar{\mathbf{f}}_{.\ell} = 0$:

$$\begin{aligned} GMD_\nu(\mathbf{a}_{.\ell}, \mathbf{z}_{.k}) &= -\frac{2\nu}{n(n-1)} \sum_{i=1}^{n} (a_{i\ell} - \bar{\mathbf{a}}_{.\ell})(r_{ik}^{\nu-1} - \bar{\mathbf{R}}_{.k}^{\nu-1}) \\ &= -\frac{2\nu}{n(n-1)} \frac{\sum_{i=1}^{n} (f_{i\ell} - \bar{\mathbf{f}}_{.\ell})(r_{ik}^{\nu-1} - \bar{\mathbf{R}}_{.k}^{\nu-1})}{GMD_\nu(\mathbf{f}_{.\ell})} \\ &= \frac{GMD_\nu(\mathbf{f}_{.\ell}, \mathbf{z}_{.k})}{GMD_\nu(\mathbf{f}_{.\ell})} \\ &= GC_\nu(\mathbf{f}_{.\ell}, \mathbf{z}_{.k}). \end{aligned}$$

Therefore, the analysis of the variability between $\mathbf{a}_{.\ell}$ and $\mathbf{z}_{.k}$ is simply the study of the correlation between $\mathbf{f}_{.\ell}$ and $\mathbf{z}_{.k}$:

$$\boldsymbol{V} = [GMD_\nu(\mathbf{a}_{.\ell}, \mathbf{z}_{.k})] = [GC_\nu(\mathbf{f}_{.\ell}, \mathbf{z}_{.k})].$$

BIBLIOGRAPHIE

Abramowitz, M. & I. Stegun. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series no. **55**.

A. Baccini, P. Besse and A. de Falguerolles (1996), A L_1 norm PCA and a heuristic approach, in *Ordinal and Symbolic Data Analysis*, E Didday, Y. Lechevalier and O. Opitz (eds), Springer, 359-368.

Banerjee, A.K. (2010), A multidimensional Gini index, *Mathematical Social Sciences*, **60**, 87-93.

Calò, D.G. (2006), On a Transvariation Based Measure of Group Separability, *Journal of Classification*, **23(1)**, 143-167.

Carcea, M. & R. Serfling (2015), A Gini autocovariance function for time series modeling. *Journal of Time Series Analysis* **36**, 817–38.

Charpentier, A., Ka, N., Mussard, S. & Ndiaye, O. (2019), Gini regressions and heteroskedasticity. *Econometrics*, **7(1)**, 4, 1-16.

Charpentier, A., Mussard, S. & Ouraga, T. (2019), Principal Componenet Analysis : A Generalized Gini Appraoch. *Working paper # 2019-02 CHROME, University of nîmes*, hal-02340386v1.

Dagum, C. (1997), A New Approach to the Decomposition of the Gini Income Inequality Ratio, *Empirical Economics*, **22**, 515-531.

Furman, E. & R. Zitikis (2017), Beyond the Pearson Correlation : Heavy-Tailed Risks, Weighted Gini Correlations, and A Gini-Type Weighted Insurance Pricing Model, *ASTIn Bulletin : The Journal of the International Actuarial Association*, **47(03)**, 919-942.

Ka, N. & Mussard, S. (2016). ℓ_1 Regressions : Gini Estimators for Fixed Effects Panel Data. *Journal of Applied Statistics*, **43(8)**, 1436-1446.

Kwak, N. Principal Component Analysis Based on L1-norm Maximization, *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30(9)**, 1672-1680.

Li, C., Shaoa, Y.-H., & Deng, N.-Y. (2015) Robust L1-norm two-dimensional linear discriminant analysis, *Neural networks*, **65**, 92-104.

Olkin, Ingram, and Shlomo Yitzhaki. 1992. Gini regression analysis. *International Statistical Review*, **60(2)**, 185-196.

Montanari, A. (2004), Linear discriminant analysis and transvariation, *Journal of Classification*, **21(1)**, 71-88.

Schechtman, E. & S. Yitzhaki (1987), A Measure of Association Based on Gini's Mean Difference, *Communications in Statistics : A*, **16**, 207–231.

Schechtman, E. and S. Yitzhaki (2003), A family of correlation coefficients based on the extended Gini index, *Journal of Economic Inequality*, **1(2)**, 129-146.

Shelef, A. (2016), A Gini-based unit root test, *Computational Statistics & Data Analysis*, **100**, 763-772.

Yaari, M.E. (1987), The Dual Theory of Choice Under Risk, *Econometrica*, **55**, 99–115.

Yaari, M.E. (1988), A Controversial Proposal Concerning Inequality Measurement, *Journal of Economic Theory*, **44**, 381–397.

Yitzhaki, S. (1994), Economic Distance and Overlapping of Distributions, *Journal of Econometrics*, **61**, 147-159.

Yitzhaki, S. (2003), Gini's Mean difference : a superior measure of variability for non-normal distributions, *Metron*, **LXI(2)**, 285-316.

Yitzhaki, S. and R. Lerman (1991), Income stratification and income inequality, *Review of Income and Wealth* **37**, 313-329.

Yitzhaki, S. & E. Schechtman (2013), *The Gini Methodology. A Primer on a Statistical Methodology*, Springer.

CONLUSION GÉNÉRALE

Dans cette thèse, nous soutenons qu'il est possible de modéliser correctement le risque en exploitant les données financières volumineuses à partir d'un cadre d'analyse construit avec des éléments issus de l'indice de Gini. L'exploitation des données volumineuses proposée ici est appliquée à la finance quantitative et couvre des analyses aussi essentielles que l'évaluation d'actifs, la mesure du risque et de performance et l'allocation optimale de portefeuille.

Depuis la fin des années 2000, les données massives (ou Big Data), particulièrement en finance constituent une source d'informations importante pour la compréhension et l'analyse des marchés financiers grâce à la modélisation des outils d'aide à la prise de décision. Elles disposent de certaines spécificités telles que la présence de valeurs extrêmes, d'outliers ou d'évènements rares qui nécessitent l'utilisation d'outils mathématiques ou économétriques robustes pour la calibration des modèles traditionnels de la gestion de portefeuilles. Ces spécificités proviennent du caractère aléatoire des cours observés et de la volatilité des marchés financiers. La présence de valeurs extrêmes, d'outliers, d'évènements rares dans les données disponibles rend préjudiciable l'estimation de nombreux modèles de la théorie financière qui sont pour la plupart linéaires. Ces particularités de l'information financière disponible ont pour conséquence le biais et la non convergence de l'estimation des risques systématiques par la méthode des moindres carrés ordinaires (Basset et Koener, 1978 ; Tanaka et al., 1982 ; Yitzhaki, 1992). Elles ont aussi pour effet de rendre moins robustes les outils statistiques de réduction de dimensions classiques que sont l'analyse en composantes principales (ACP) et l'analyse linéaire discriminante (LDA) utilisés pour la sélection d'actifs par la méthode d'arbitrage et la prédiction de portefeuilles "verts" (portefeuilles d'actifs issus de titres financiers de sociétés ayant une gouvernance responsable). Shalit et Yitzhaki (1989) montrent l'efficacité de la sélection de portefeuilles d'actifs par une approche Gini en utilisant la GMD comme mesure de risque. De plus l'utilité de l'opérateur coGini en analyse du risque a fait ses preuves comme le montrent Furman & Zitikis (2017). Banerjee (2010) ouvre une voie à l'analyse statistique multidimensionnelle avec une approche Gini en proposant un indice de Gini multidimensionnel qui satisfait simultanément la condition de la majorisation croissante de la corrélation et la condition de la comonotonicité unidirectionnelle.

Le volet théorique de cette thèse concerne la mise en place d'un cadre mathématique et statistique pour le développement d'outils et méthodes robustes. Il peut être perçu comme suit :

La démarche théorique dans un premier temps s'appuie principalement sur les travaux de Banerjee (2010) et sur la décomposition de la matrice de corrélation au sens de Gini pour proposer des outils statistiques multidimensionnelles à l'aide de l'indice de Gini. Dans un second temps, elle met en exergue l'impact des observations extrêmes présentes dans les données sur la fonction quadratique communément utilisée en finance (la variance) à travers une revue de littérature sur la régression Gini.

Le premier cadre d'analyse met en évidence des outils d'analyse factorielle plus robustes aux outliers que les outils conventionnels. Il s'agit en l'occurence de l'analyse en composantes principales (ACP) et de l'analyse linéaire discriminante (ALD ou AFD). En lieu et place de la matrice de corrélation au sens de Pearson, celle au sens du Gini est maximisée pour obtenir dans le cadre l'ACP les composantes principales et dans celui le l'AFD des facteurs discriminants, *C.f* Ouraga (2019) et Candevaux et al. (2020). Ces outils servent ensuite de support à des modèles de la théorie financière. Le second volet de cette démarche théorique est fait pour revisiter des classiques de la théorie financière dont l'évaluation d'actif par la méthode d'arbitrage avec des facteurs communs de risque issus de l'ACP-Gini, de la frontière efficiente, des ratios de performance. Cet outil de gestion de portefeuilles est fortement tributaire des estimateurs du risque systématique. Nous avons vu dans le chapitre 1, les conséquences des valeurs extrêmes sur l'estimation de celui-ci par les MCO qui est le modèle économétrique couramment utilisé en finance pour l'APT classique.

Les apports théoriques importants des outils proposés par des simulations de Monte Carlo :

— La mise en place de Théorèmes sur des approches importantes de la théorie financière avec une approche Moyenne-GMD : la diversification, la frontière efficiente et le portefeuille de marché.
— La robustesse de l'approche Gini des analyses discriminantes linéaires et quadratiques.

— La robustesse de l'ACP-Gini (simple et généralisée).
— La robustesse de l'APT-Gini : stabilité des primes de risques.
— La robustesse du ratio de Treynor généralisé(RTG-Gini).

Les apports empiriques des outils proposés lors de la démarche sont appliqués à des données financières journalières. Ces études empiriques conduisent aux résultats suivants :

— L'application du modèle d'évaluation par arbitrage à facteurs latents sur 8 actifs du marché financier français et 17 facteurs sur la période allant du 03/01/2008 au 30/12/2010 reflète la situation des marchés financiers sur ladite période (crise des subprimes). Les primes de risque unitaire par l'approche Gini sont plus faibles que celles de la méthode classique mais stables contrairement à celles de l'APT classique.
— L'application de scoring aux portefeuilles "verts" du CAC-40 sur la période allant du 30/01/2018 au 01/02/2019 confirmes les prédictions théoriques, à savoir que l'algorithme de prédiction proposé est meilleur en présence d'individus extrêmes avec un taux d'erreur global qui est nul.

Cette thèse permet ainsi de proposer de nouveaux outils d'exploitation de l'information dans les données massives en générale et dans les données financières pour la gestion de portefeuilles en particulier. Précisons que la variance peut être remplacée par la GMD qui utilise le rang des observations pour réduire sensiblement l'influence des observations extrêmes. Par ces résultats, nous élargissons le champ des possibles au niveau des modèles classiques de la théorie financière en tenant compte de l'influence des outliers qui caractérisent les marchés financiers. Au-delà de la finance, l'on élargit également la portée des modèles de réduction d'espace et de classification basés sur les métriques Gini. Les analyses théoriques ont mis en avant l'apport de cette thèse à la fois à la gestion de portefeuille et à l'analyse de données.

Les outils et méthodes d'analyses proposés dans cette thèse peuvent être améliorés. Nous présentons quelques axes de recherche comme perspectives pour les travaux présentés dans cette thèse.

L'utilisation du GMD dans des modèles dynamiques. Il serait intéressant d'avoir des modèles d'allocation dynamiques, qui considèrent donc la possibilité d'investir sur plusieurs périodes avec le GMD comme mesure de risque.

L'application à des données financières à haute fréquence. Les données financières à haute fréquence donnent accès à des volumes de données plus importantes et la présence éventuelle de plus d'observations extrêmes. Il existe cependant du bruit microstrucurel qui accompagne les données à haute fréquence, qui est dû à la fréquence très courte de l'enregistrement des données. Mbairadjim (2014) dans sa thèse propose d'utiliser la rentabilité floue des actifs pour l'évaluation des modèles. Nous évoluerons par conséquent dans un modèle espérance de rentabilité floue - GMD.

L'introduction des moments d'ordre supérieurs. Il est nécessaire de développer des moments centrés d'ordre 3 (skewness) et d'ordre 4 (kurtosis) au sens de Gini pour l'allocation optimale de portefeuilles mais aussi dans les modèles d'évaluation d'actifs et dans les mesures de risques.

L'extension de l'analyse de données au sens de Gini. Il a été mis en évidence dans ces travaux de thèse la robustesse de l'ACP et de l'ALD en présence d'observations extrêmes. La continuité logique serait de l'étendre à d'autres méthodes de l'analyse classique des données.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX