Principal Component Analysis: A Generalized Gini Approach

Arthur Charpentier, Stéphane Mussard, Tea Ouraga

Principal Component Analysis: A Generalized Gini Approach

Arthur Charpentier
UQAM

Stéphane Mussard*
Chrome
Université de Nîmes

Téa Ouraga
Chrome
Université de Nîmes

Abstract

A principal component analysis based on the generalized Gini correlation index is proposed (Gini PCA). The Gini PCA generalizes the standard PCA based on the variance. 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.

keywords : Generalized Gini ; PCA ; Robustness.

*Université de Nîmes – e-mail: stephane.mussard@unimes.fr, Research fellow Grédi University of Sherbrooke, and Liser Luxembourg.

1 Introduction

This late decade, a line of research has been developed and focused on the Gini methodology, see Yitzhaki & Schechtman (2013) 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 paper, 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 constant when the data are transformed by monotonic maps.[2] Another property typically related to the

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

[2] See Furman & Zitikis (2017) for the link between variability (risk) measures and 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).

In the field of PCA, Baccini, Besse & de Falguerolles (1996) and Korhonen & Siljamäki (1998) are 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 metric introduced by Gini (1912), which is a city-block distance measure of variability. The authors 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 the same vein, Ding *et al.* (2006) make use of a rotational ℓ_1 norm PCA to robustify the variance-covariance matrix in such a way that the PCA is rotational invariant. 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).

In this paper, 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

Gini correlation index.

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 2 sets the notations and presents some ℓ_2 norm approaches of PCA. Section 3 reviews the Gini-covariance operator. Section 4 is devoted to the generalized Gini PCA. Section 5 focuses on the interpretation of the Gini PCA. Sections 6 and 7 present some Monte Carlo simulations and applications, respectively.

2 Motivations for the use of Gini PCA

In this Section, the notations are set. Then, some assumptions are imposed and some ℓ_2-norm PCA techniques are reviewed in order to motivate the employ of the Gini PCA.

2.1 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 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.

2.2 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

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

or equivalently

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

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

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

With scaled variables[3] (i.e. $\operatorname{Var}[\mathbf{x}_{\cdot,j}] = 1$, $\forall j$)

$$\omega_1^* \in \operatorname{argmax} \left\{ \sum_{j=1}^{K} \operatorname{Cor}[\mathbf{x}_{\cdot,j}, \boldsymbol{X}\omega] \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1. \tag{3}$$

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

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

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 $\operatorname{Var}[\boldsymbol{X}\omega]$ 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

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

[3] 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 $\mathbf{\Sigma}_{(k-1)} = \mathbf{\Sigma}_{(k-2)} - \omega_{k-1}^* \omega_{k-1}^{*\top} \mathbf{\Sigma}_{(k-1)} \omega_{k-1}^* \omega_{k-1}^{*\top}$ (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)[4].

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

$$\omega_1^* \in \text{argmax} \left\{ \sum_{j=1}^{K} \mathcal{R}[\mathbf{x}_{\cdot,j}, \boldsymbol{X}\omega] \right\} \text{ subject to } \|\omega\|_2^2 = \omega^\top \omega = 1 \tag{4}$$

where $\mathcal{R}$ denotes some rank based correlation, *e.g.* Spearman's rank correlation, as an extention of Equation (3). So, quite naturally, one possible

[4]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 (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 (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 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.

3.1 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,

$$\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^\intercal]. \tag{5}$$

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 3.1. *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 3. Nevertheless, a stronger technical assumption will be necessary: pairwise-exchangeability.

Definition 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 3.1. *If* $\boldsymbol{X}$ *is a Gaussian vector with identically distributed components, then* $\boldsymbol{X}$ *is pairwise-exchangeable.*

Proof. 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:

$$\Delta(X) = \mathbb{E}(|X_1 - X_2|) \text{ where } X_1, X_2 \sim X, \text{ and } X_1 \perp\!\!\!\perp X_2, \tag{6}$$

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. *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 $\Gamma C_F(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{X}, F(\boldsymbol{X}))$.*

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

Theorem 3.1. *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) $\Gamma C_F(\boldsymbol{X}) = \sigma^{-1}\mathrm{Cov}(Z, \Phi(Z))\mathrm{Var}(\boldsymbol{X})$.

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

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

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

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

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

which ends the proof. □

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

Definition 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 $G\Gamma C_\nu(\boldsymbol{X}) = \Gamma C_{-\overline{F}^{\nu-1}}(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{X}, -\overline{F}^{\nu-1}(\boldsymbol{X}))$, 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:

$$GMD_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k}) := -\frac{2}{N-1}\nu \text{Cov}(\mathbf{x}_{\cdot\ell}, \mathbf{r}_{\mathbf{x}_{\cdot k}}^{\nu-1}),\ \nu > 1,$$

where $\mathbf{r}_{\mathbf{x}_{\cdot k}} = (R(x_{1k}), \ldots, R(x_{nk}))$ is the decumulative rank vector of $\mathbf{x}_{\cdot k}$, 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 $GMD_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k})$ is the empirical version of

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

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

$$GMD_2(X_k, X_k) = 4\text{Cov}\big(X_k, F_k(X_k)\big) = \Delta(X_k).$$

When $k = \ell$, GMD_ν represents the variability of the variable $\mathbf{x}_{\cdot k}$ itself. Focus is put on the lower tail of the distribution $\mathbf{x}_{\cdot k}$ 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 $\mathbf{x}_{\cdot k}$. 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.[5]

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 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 $\mathbf{x}_{\cdot k}$ being not a monotonic transformation of $\mathbf{x}_{\cdot\ell}$ and $\nu > 1$, $GMD_\nu(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot\ell}) \neq GMD_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot k})$.

[5]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).

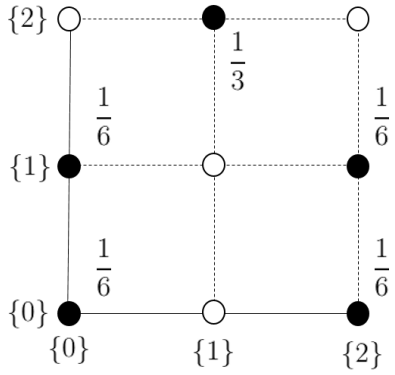


Figure 1: Joint distribution of a random pair (X_k, X_ℓ) such that $\mathbb{E}[X_k h(X_\ell)] \neq \mathbb{E}[X_\ell h(X_k)]$, with non-exchangeable components $X_k \overset{\mathcal{L}}{=} X_\ell$.

3.2 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 3.1. – **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 3.1 (iv), it is possible to rescale the variables $\mathbf{x}_{\cdot\ell}$ thanks to a linear transformation, then the matrix of standardized observation is,

$$\boldsymbol{Z} \equiv [z_{i\ell}] := \left[\frac{x_{i\ell} - \bar{\mathbf{x}}_{\cdot\ell}}{GMD_\nu(\mathbf{x}_{\cdot\ell}, \mathbf{x}_{\cdot\ell})}\right]. \tag{7}$$

The variable $z_{i\ell}$ is a real number without dimension. The variables $\mathbf{x}_{\cdot k}$ 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 $\boldsymbol{X}$:

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

Note that the last equality holds since the standardization (7) is a strictly increasing affine transformation.[6] The $K \times K$ matrix containing all G-correlation indices between all couples of variables $\mathbf{z}_{\cdot k}$ and $\mathbf{z}_{\cdot\ell}$, 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 $GMD_\nu(\boldsymbol{Z}) \equiv [GMD_\nu(\mathbf{z}_{\cdot k}, \mathbf{z}_{\cdot\ell})]$, then we get the following.

Proposition 3.2. *For each standardized matrix $\boldsymbol{Z}$ defined in* (7)*, the following relations hold:*

$$GMD_\nu(\boldsymbol{Z}) = GC_\nu(\boldsymbol{X}) = GC_\nu(\boldsymbol{Z}). \tag{8}$$

$$GMD_\nu(\mathbf{z}_{\cdot k}, \mathbf{z}_{\cdot k}) = 1, \ \forall k = 1, \ldots, K. \tag{9}$$

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

Proof. 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,

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

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

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

By Property 3.2 (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 3.2. *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.*

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

Theorem 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 4 that Gini PCAs and classical PCA (based on the ℓ_2 norm and the covariance matrix) are equivalent.

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 $\{\mathbf{z}_{.1}, \ldots, \mathbf{z}_{.K}\}$. 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.

4.1 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 $\mathbf{f}_{.k}$ 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 $\boldsymbol{F} \equiv [\mathbf{f}_{.1}, \ldots, \mathbf{f}_{.K}]$ with $\boldsymbol{R}^c_{\mathbf{f}} \equiv [\mathbf{r}^{\nu-1}_{c,\mathbf{f}_1}, \ldots, \mathbf{r}^{c}_{c,\mathbf{f}_{.K}}]$ its corresponding decumulative centered rank matrix (where each decumulative rank vector is raised to an exponent of $\nu - 1$). The $K \times K$ matrix $\boldsymbol{B} \equiv [\mathbf{b}_{.1}, \ldots, \mathbf{b}_{.K}]$ is the projector of the observations, with the normalization condition $\mathbf{b}_{.k}^\top \mathbf{b}_{.k} = 1$, such that $\boldsymbol{F} = \boldsymbol{Z}\boldsymbol{B}$. We denote by $\lambda_{.k}$ (or $2\mu_{.k}$) the eigenvalues of the matrix $[GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top]$. Let the basis $\mathscr{B} := \{\mathbf{b}_{.1}, \ldots, \mathbf{b}_{.h}\}$ with $h \leq K$ issued from the maximization of the overall Gini variability:

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

Indeed, from the Lagrangian,

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

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

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

that is,

$$[GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top]\mathbf{b}_{\cdot k} = \lambda_{\cdot k}\mathbf{b}_{\cdot k}. \tag{10}$$

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 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* $\mathbf{r}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} = \boldsymbol{R}_{\mathbf{z}}^c\mathbf{b}_{\cdot k}$.
(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* $k = 1, \ldots, K$.

Proof. (i) Note that $GMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = -\frac{2\nu}{N(N-1)}(\mathbf{f}_{\cdot k} - \bar{\mathbf{f}})^\top \mathbf{r}_{c,\mathbf{f}_{\cdot k}}^{\nu-1}$, where $\mathbf{r}_{c,\mathbf{f}_{\cdot k}}^{\nu-1}$ 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:[7]

$$\begin{aligned}
\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k} &= -\frac{2\nu}{N(N-1)}\mathbf{b}_{\cdot k}^\top \boldsymbol{Z}^\top \boldsymbol{R}_{\mathbf{z}}^c\mathbf{b}_{\cdot k} \\
&= -\frac{2\nu}{N(N-1)}\mathbf{b}_{\cdot k}^\top \boldsymbol{Z}^\top \mathbf{r}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} \quad (\text{by } \mathbf{r}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} = \boldsymbol{R}_{\mathbf{z}}^c\mathbf{b}_{\cdot k}) \\
&= -\frac{2\nu}{N(N-1)}\mathbf{f}_{\cdot k}^\top \mathbf{r}_{c,\mathbf{f}_{\cdot k}}^{\nu-1} \\
&= GMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}).
\end{aligned} \tag{11}$$

Then, maximizing the multidimensional variability $\mathbf{b}_{\cdot k}^\top GC_\nu(\boldsymbol{Z})\mathbf{b}_{\cdot k}$ yields from (10):

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

[7] 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. □

4.2 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).[8]

[8]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 $GGMD_\nu$.

4.3 Properties of $GGMD_\nu$

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

Proposition 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) $\min_{k=1,\ldots,K} GGMD_\nu(\mathbf{f}_{\cdot k}, \mathbf{f}_{\cdot k}) = \mu_{\cdot K}$.

Proof. (i) The result comes from the rank-nullity theorem. From the eigenvalue Equation (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 $\mathbf{r}^c_{\mathbf{f}_{\cdot k}} = \mathbf{0}$, 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:

$$\lambda_{\max} := \max \frac{\mathbf{b}_{\cdot 1}^\top [GC_\nu(\boldsymbol{Z}) + GC_\nu(\boldsymbol{Z})^\top] \mathbf{b}_1}{\mathbf{b}_1^\top \mathbf{b}_1} = \lambda_1.$$

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 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* $\nu > 1$*:*

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

Proof. From Theorem 3.2:

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

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

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

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:

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

4.4 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.

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.

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

so

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

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

5.2 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 $\mathbf{f}_{\cdot k}$. Let us denote $\widetilde{U}_{\ell k} := \text{Cov}(\mathbf{f}_{\cdot \ell}, \boldsymbol{R}^c_{\mathbf{z}_{\cdot k}})$ with $\boldsymbol{R}^c_{\mathbf{z}_{\cdot k}}$ the (decumulative) centered rank vector of $\mathbf{z}_{\cdot k}$ raised to an exponent of $\nu - 1$ and $U_{\cdot \ell} := \text{Cov}(\mathbf{f}_{\cdot \ell}, \boldsymbol{R}^c_{\mathbf{f}_{\cdot \ell}})$. Those two Gini covariances yield the following U-statistics:

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

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:[9]

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

The usual PCA enables the variables to be analyzed in the circle of correlation, which outlines the correlations between the variables $\mathbf{z}_{\cdot k}$ and the

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

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 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.

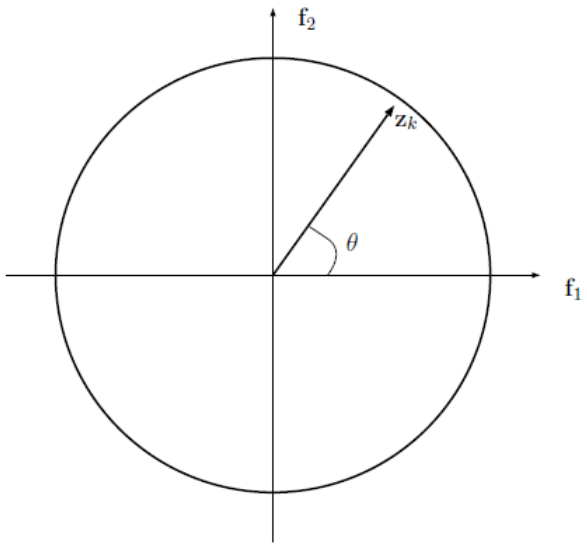


Figure 2: Circle of correlation

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.

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 1: 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:

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

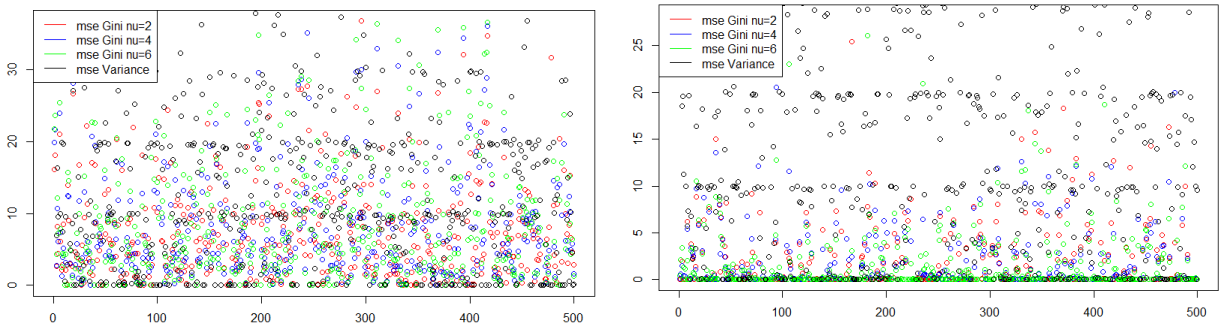
As can be seen in the matrix above, we can expect that all the information be gathered on the first axis because each pair of variables records an important linear correlation. The repartition of the information on each component, that is, each eigenvalue in percentage of the sum of the eigenvalues is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

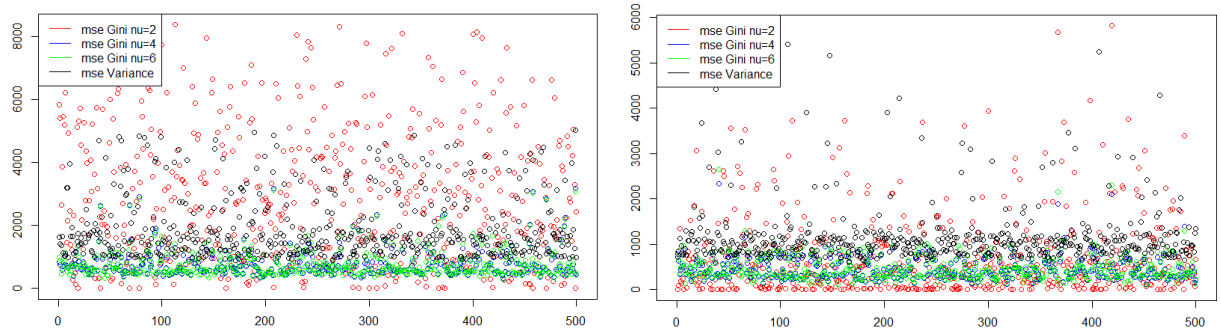
Table 1: Eigenvalues and their MSE

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 3a-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 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 2).



(a) ACT_1 (b) ACT_2



(c) RCT_1 (d) RCT_2

Figure 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 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.

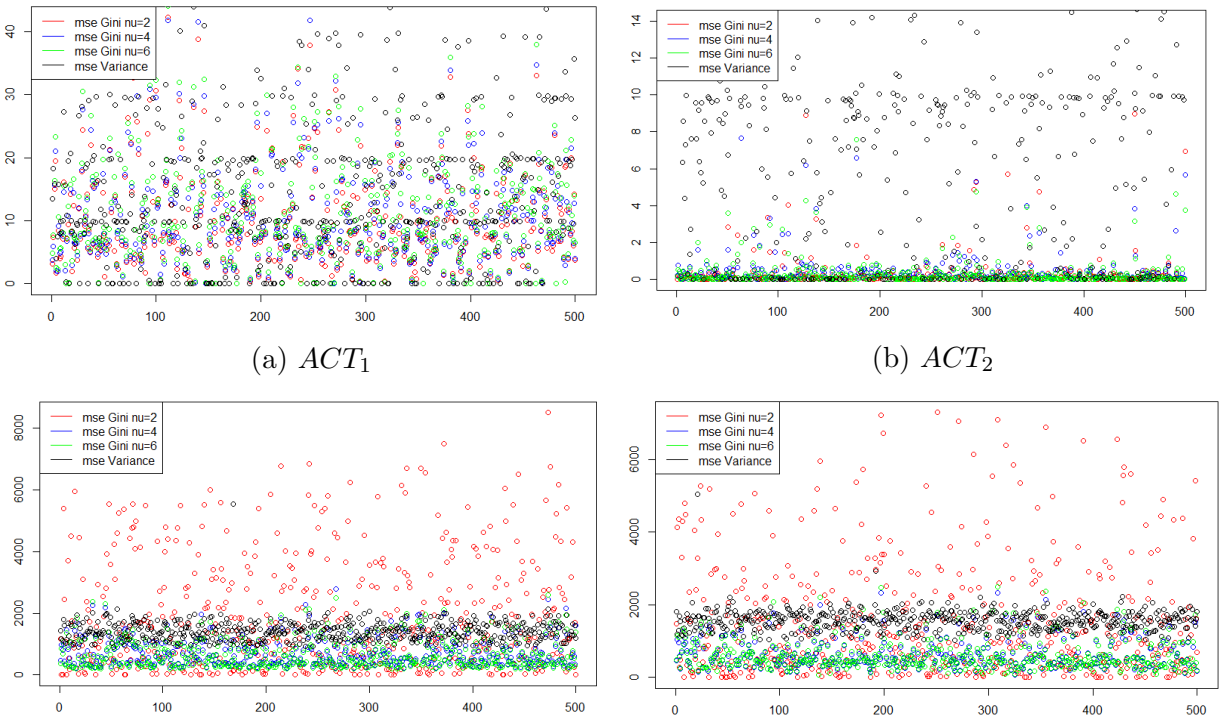
$$\rho = \begin{pmatrix} 1 & -0.5 & 0.25 & 0.5 \\ -0.5 & 1 & -0.9 & 1 \\ 0.25 & -0.9 & 1 & -0.25 \\ 0.5 & 0 & -0.25 & 1 \end{pmatrix}$$

The repartition of the information over the new axes (percentage of each eigenvalue) is given in Table 3. 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: 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 (4a-4b). We obtain the same kind of results, with less variability on the second axis. In Figures 4a-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 4c-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 4: ACT_1, ACT_2, RCT_1 and RCT_2

The results of Figures 4a-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 4: Standard deviation of the MSE of the ACT on the two first axes

As in the previous example of simulation, Table 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.

7 Application on cars data

We propose a simple application with the celebrated cars data (see the Appendix).[10] The dataset is particularly interesting since there are highly correlated variables as can be seen in the Pearson correlation matrix given in Table 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 5: Correlation matrix

Also, the dataset is composed of some outlying observations (Figure 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 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 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 that

[10] 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

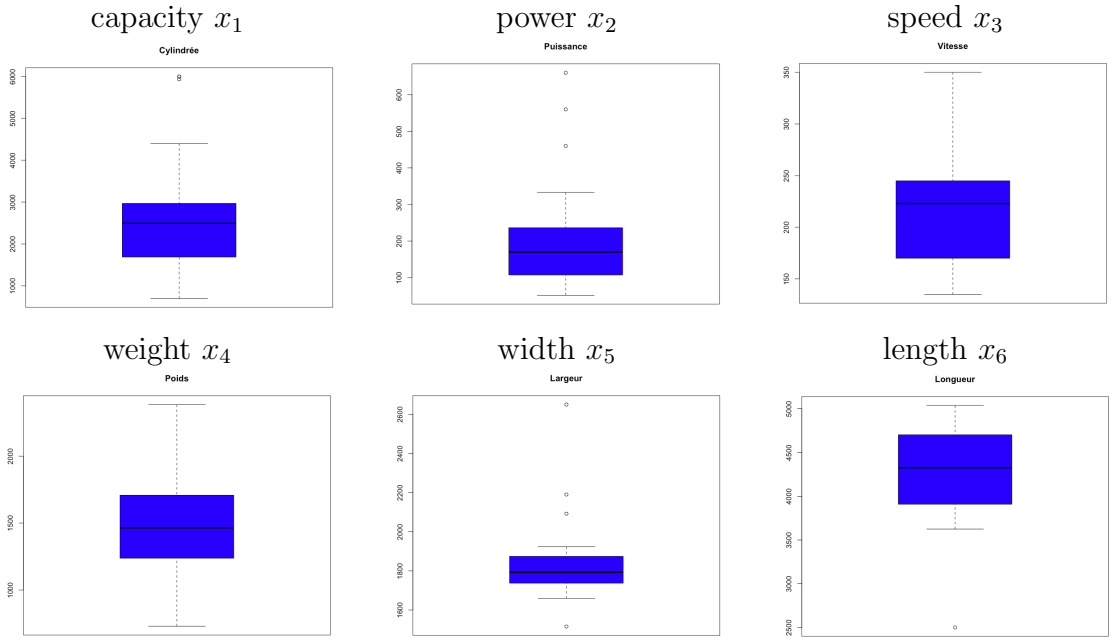
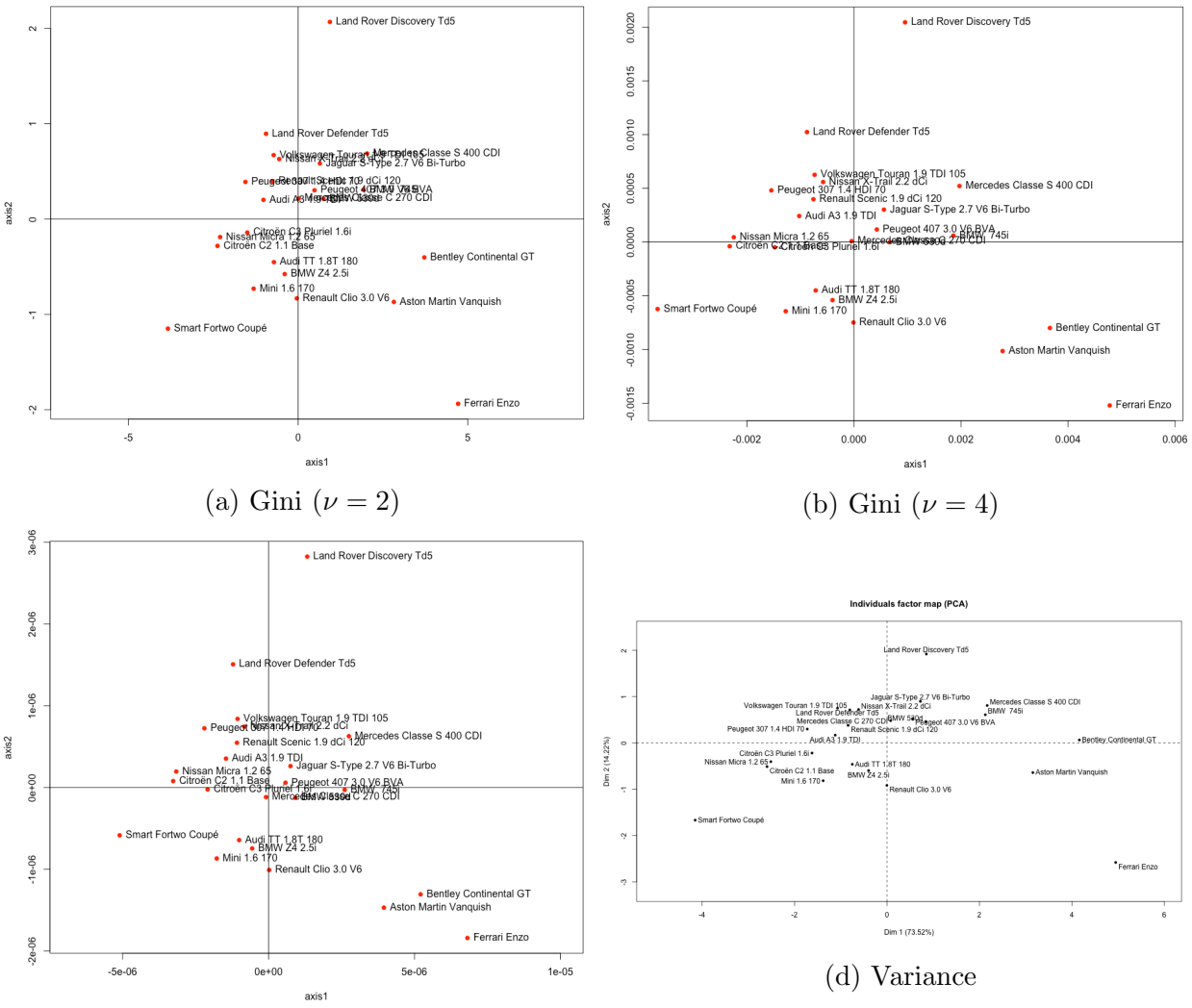


Figure 5: Box plots

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 6, for each method.

As depicted in Figure 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 to interpret the



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

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

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

(d) Variance

Figure 6: Projections of the cars

results, see Tables 7 to 10.

Some slight differences appear between the Gini PCA and the classical one based on the variance. The theoretical Section 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 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 8: Correlations Axes / variables (significance 5%)

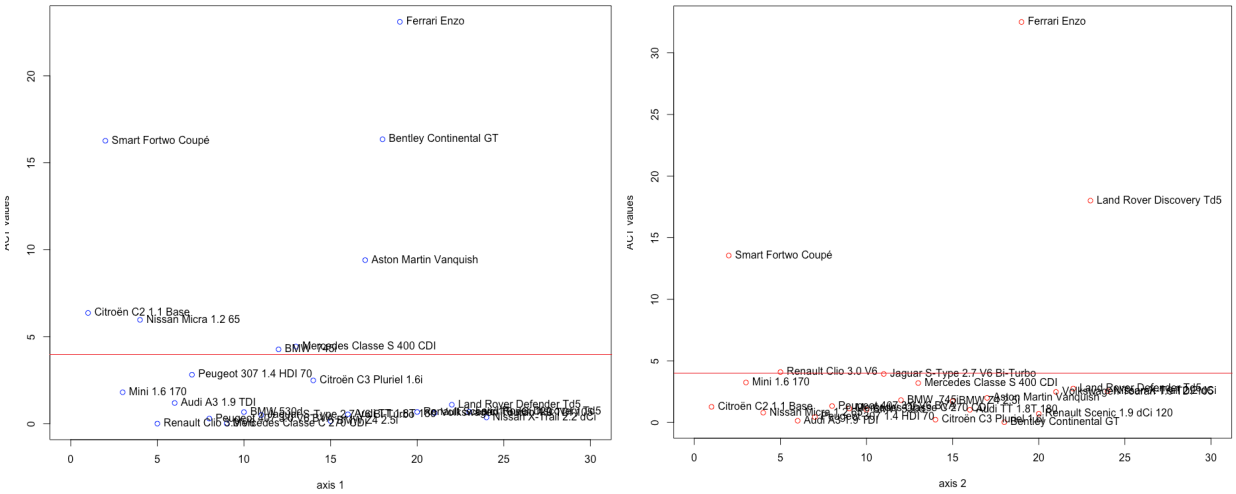
Gini ($\nu = 6$)		capacity	power	speed	weight	width	length
Axe 1	valeurs	-0.781	-0.759	-0.598	-0.730	-0.755	-0.701
	U**-stat**	**-4.036**	**-3.903**	**-3.137**	**-3.125**	**-3.882**	**-3.644**
Axe 2	valeurs	0.019	-0.170	-0.570	0.153	0.125	0.218
	U**-stat**	**0.089**	**-0.734**	**-1.914**	**0.734**	**0.569**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

Table 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 10: Correlations Axes / variables (significance 5%, 10%)

Tables 8 and 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 8, 9, 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.

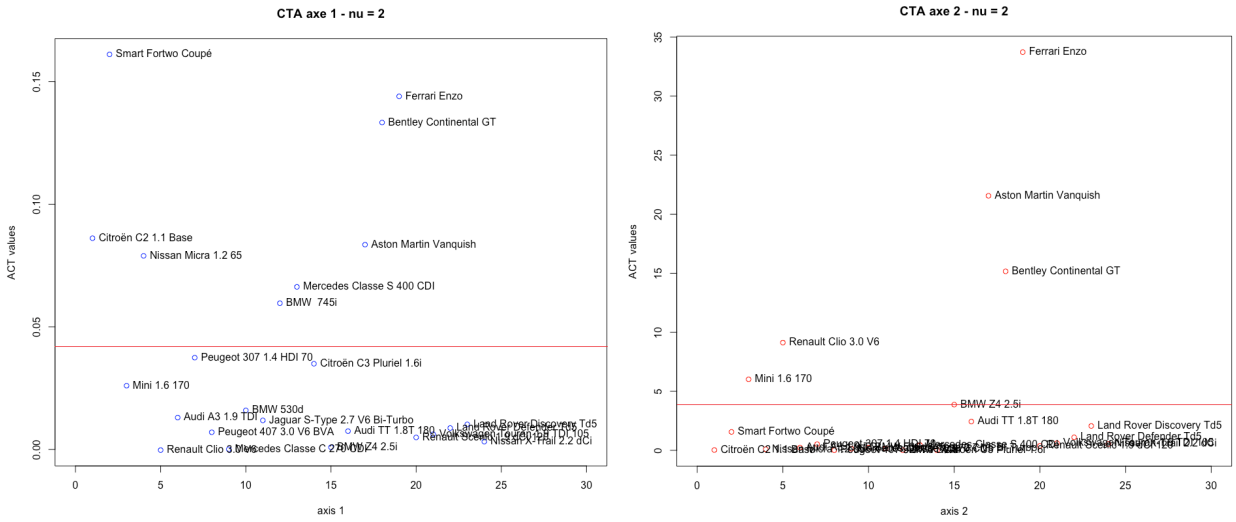


(a) Axis 1 (variance)

(b) Axis 2 (variance)

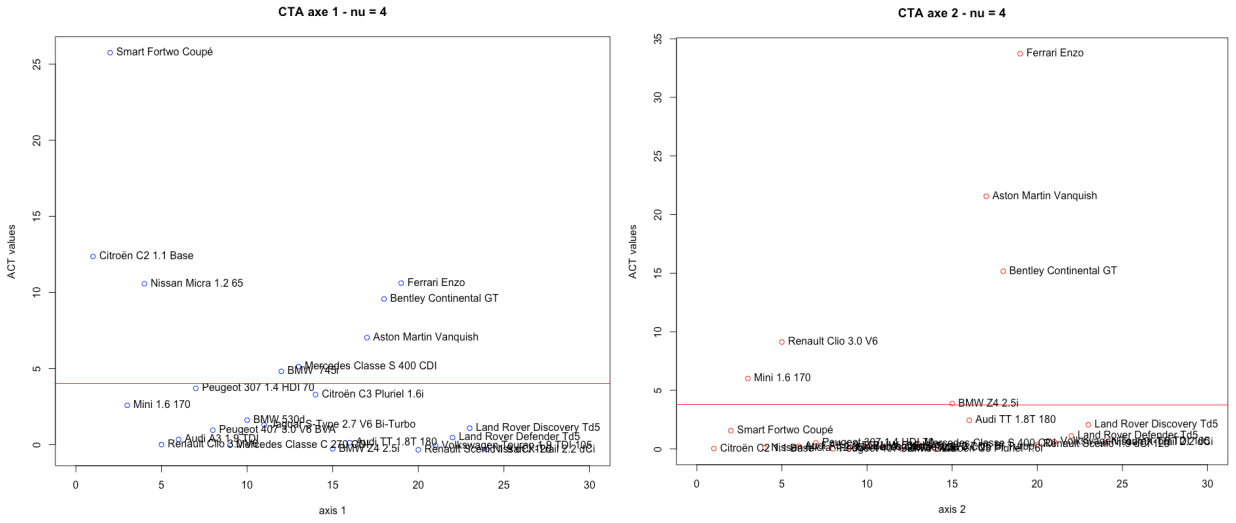
Figure 7: Variance ACTs

Finally, some cars are not correlated with axis 2 in the standard PCA, see Figures 8–10, while this is the case in the Gini PCA. Indeed some cars are



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

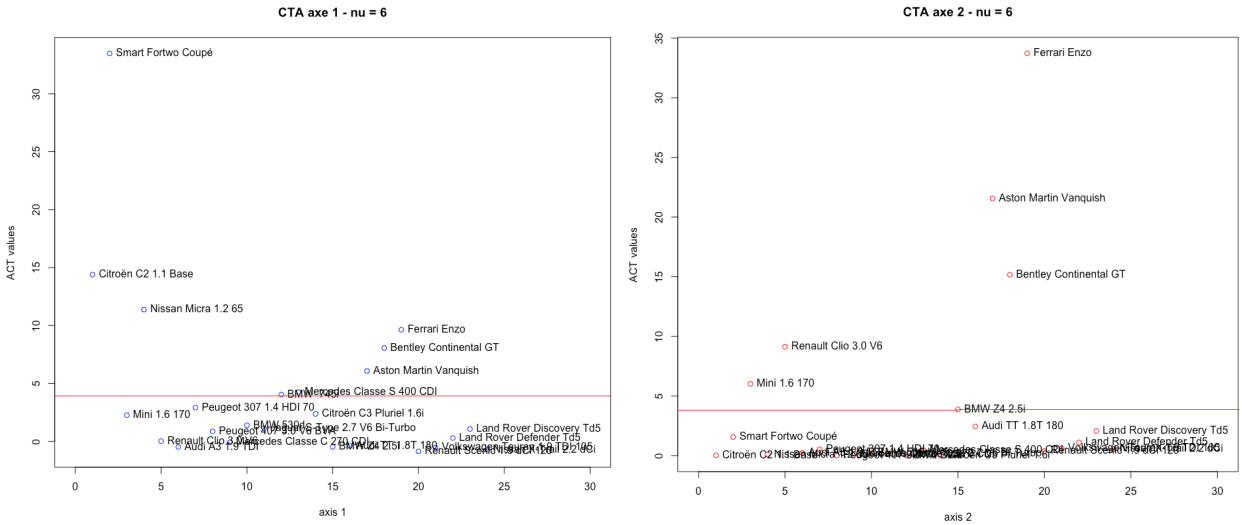
Figure 8: Gini ACTs ($\nu = 2$)



(a) Axis 1 ($\nu = 4$) (b) Axis 2 ($\nu = 4$)

Figure 9: Gini ACTs ($\nu = 4$)

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



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

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

Figure 10: Gini ACTs ($\nu = 6$)

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

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.

References

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.

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.

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.

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.

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.

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 11: Cars data