LIS
Working Paper Series

No. 651

Comparing the homogeneity of income distributions using Polarization Indices

André-Marie Taptué

September 2015

Luxembourg Income Study (LIS), asbl

Comparing the homogeneity of income distributions using Polarization Indices

André-Marie TAPTUÉ*

Abstract

In the context of polarized societies, income homogeneity is linked to the frequency and the intensity of social unrest. Most homogenous countries exhibit a lower frequency of intense social conflicts and less homogeneous countries show a higher frequency of moderate social conflicts. This paper develops a methodology to compare the degree of homogeneity of two income distributions. We use for that purpose an index of polarization that does not account for alienation. This index is the identification component of polarization that measures the degree to which individuals feel alike in an income distribution. This development leads to identification dominance curves and derives first-order and higher-order stochastic dominance conditions. First-order stochastic dominance is performed trough identification dominance curves drawn on a support of identification thresholds. These curves are used to determine whether identification, homogeneity, or similarity of individuals is greater in one distribution than in another for general classes of polarization indices and ranges of possible identification thresholds. We also derive the asymptotic sampling distribution of identification dominance curves and test dominance between two distributions using Intersection Union tests and bootstrapped p-values. Our methodology is illustrated by comparing pairs of distributions of eleven countries drawn from the Luxembourg Income Study database.

Key words: Alienation, identification, polarization, stochastic dominance.

*Departement D'économique and CIRPÉE, Université Laval, Canada. email: andre-marie.taptue.1@ulaval.ca

1 Introduction

Income polarization is a consequence of an interaction between identification and alienation in an income distribution. Alienation measures dissimilarity between individuals with different incomes and identification measures the degree to which people belonging to the same group feel alike and identified to each other. An interaction between alienation and identification creates antagonism between groups and can lead to social unrest and civil war (Esteban and Ray, 1994). However, putting a part discrepancies between incomes, social unrest is frequent in heterogeneous countries where people feel less alike and less identified to each other and it is rare in homogeneous countries where inhabitants feel more similar to each other in terms of income and even in terms of social characteristics. Therefore, choosing an income distribution that bears more homogeneity and identification within inhabitants can reduce the risk of conflict in a society.

Polarization studies measure identification with group sizes and assume that a distribution composed of many groups with small sizes is less homogeneous than a distribution composed of few groups with large sizes, which is said to be more homogenous (Esteban and Ray, 1994). In the former distribution, the frequency of social unrest is high though of a moderate intensity, and individuals feel less alike and less identified to each other than in the latter distribution that witnesses a lower frequency of social unrest though intense when it occurs (Esteban and Ray, 2008). The best situation is when all individuals are encompassed in one income group. Hence, identification as measured in polarization can reveal not only the degree of similarity of individuals, but also the homogeneity of a distribution and the risk of social conflict.

This paper focuses on the identification component in a polarization index to order distributions by neglecting the interaction between identification and alienation. Integrating this procedure in a polarization index leads to a reduced form of the index that measures the equality-like component of polarization or the within-group homogeneity of a distribution. The equality and the homogeneity of a distribution correspond to the presence of many similar individuals with the same characteristics of interest such as income. Using the reduced form of a polarization index to compare homogeneity of two income distributions means to compare the importance of group densities in the two distributions.

Our methodology implements stochastic dominance technics to allow for robust comparisons of homogeneity between two distributions. With stochastic dominance, we go across a range of identification thresholds that represent densities that must not be exceeded by a density of any income group and derive an identification dominance curve. First-order stochastic dominance may hold until a certain identification threshold lower than some acceptable values. The analysis can then be extended to second-order stochastic dominance and then to higher orders. On the whole, the exercise is performed continuously and it stops either when a maximum threshold of identification is reached or when the order of the dominance is judged sufficiently high. A distribution has larger identification, is more homogenous or its individuals feel more alike if its dominance curve is always above that of a second distribution.

We also develop statistical inference tests and infer stochastic dominance from data samples for robust conclusions. We derive the sampling distribution of an estimator of the stochastic dominance curves and we perform statistical inference using bootstrapped p-values.

We begin in Section 2 by showing how identification is measured in polarization indices. Section 3 develops first-order, second-order and higher-order stochastic dominance, and characterizes the classes of polarization indices compatible with each order of dominance. An identification dominance curve is developed in the particular case of income distributions in Section 4. Section 5 provides statistical inference by deriving the asymptotic properties of an estimator of dominance curves and implementing statistical tests. Section 6 compares income distributions of 11 countries using 2004 data from the Luxembourg Income Study database and Section 7 concludes.

2 Identification in polarization measures

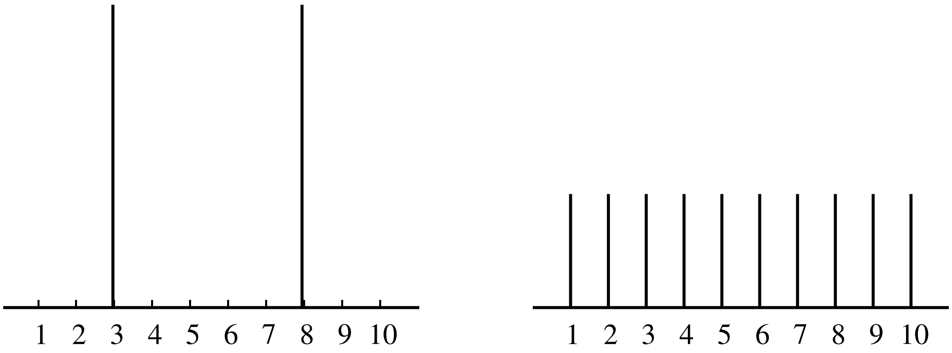
This section reviews some polarization indices in order to see how identification has been taken into account. In the framework of alienation and identification, Esteban and Ray (1994) suppose a population grouped into significantly-sized clusters such that members of each cluster are similar in terms of an attribute. The following citation shows the importance of group sizes and homogeneity in a distribution :

" *As the struggle proceeds, 'the whole society breaks up more and more into two hostile camps, two great, directly antagonistic classes: bourgeoisie and proletariat.' The classes polarize, so that they become internally more* ***homogeneous*** *and more and more sharply distinguished from one another in wealth and power* " (Esteban and Ray (1994), p. 820).

This view of homogeneity is supported by Figure 1 first built by Esteban and Ray (1994) to illustrate that there can be a high degree of within-group homogeneity in a polarized society. It shows that, departing from a society divided into two large groups relative to an attribute, group identification and consequently polarization can be reduced by clustering individuals in ten groups instead of two without caring about distances between them.

In Figure 1, the sizes of groups have decreased as well as polarization, while no attention has been paid to distances between groups during the process. For instance, consider two individuals respectively in groups 3 and 8 on the graph on the left, and suppose they are displaced respectively to groups 1 and 10 on the graph on the right. The distance between the two individuals has augmented from 5 to 9, but they are now in groups of smaller sizes and polarization has decreased. However, the risk of the emergence of social conflict with a lower intensity is higher in the situation with 10 groups (Esteban and Ray, 2008).

Figure 1: Reducing group sizes also reduces identification: A society composed of few groups of large sizes has more identification than a society composed of many groups of small sizes.



Duclos et al. (2004) use the traditional framework of alienation/identification for the measurement of polarization in the case where the distribution of the attribute of interest is described by a continuous density function. They propose the following formulation of the polarization index as the sum of all effective antagonisms in a society with a focus on income distributions:

$$P = \iint p(a(x,y),\iota(x))f(x)f(y)dxdy \tag{1}$$

where p is the antagonism function, $a(x,y)$ is the alienation function between the groups with income x and y, $\iota(x)$ is the identification in the group with income x and $f(.)$ is the income density. Then, identification in a group can be measured as the density of income in that group (Duclos et al. (2004), p. 1740):

$$\iota(x) = f(x) \tag{2}$$

In a situation of pure income polarization where individuals identify themselves only with those with similar income, group sizes are estimated non-parametrically using kernel procedures.

3 Stochastic dominance

3.1 First-order stochastic dominance

Let P_A be a polarization index in a population A as defined in Duclos et al. (2004) and written here in its general form:

$$\begin{aligned} P_A &= \iint p(a,i)dF(a,i) \qquad (3) \\ &= \int_0^{\infty}\int_0^{\infty} p(a,i)f(a,i)dadi \qquad (4) \end{aligned}$$

where $p(.,.)$ is the antagonism function taking alienation and identification as arguments, $F(.,.)$ is the joint cumulative distribution function of alienation and identification and $f(.,.)$ the corresponding joint density. Alienation and identification are two variables defined in $\mathbb{R}^+$.

When alienation is negligible in a society, the antagonism function no longer depends on it so that $p(a,i) \propto p(i)$ and we can set

$$P_A = \int_0^{\infty} p(i)f(i)di \qquad (5)$$

where $f(i)$ is the density of identification. For a decomposition of P_A, we integrate it proceeding first to a substitution: $\iota = -i$. Then,

$$P_A = \int_{-\infty}^{0} p(-\iota)f(-\iota)d\iota \qquad (6)$$

Next, integrating P_A by parts yields

$$\begin{aligned} P_A &= p(-\iota)\int_{-\infty}^{\iota} f(-i)di|_{\iota=-\infty}^{\iota=0} + \int_{-\infty}^{0} p^1(-\iota)\left(\int_{-\infty}^{\iota} f(-i)di\right)d\iota \qquad (7) \\ &= p(0)\int_{-\infty}^{0} f(-i)di + \int_{-\infty}^{0} p^1(-\iota)\left(\int_{-\infty}^{\iota} f(-i)di\right)d\iota \qquad (8) \\ &= p(0)\int_{0}^{\infty} f(i)di + \int_{0}^{\infty} p^1(\iota)\left(\int_{\iota}^{\infty} f(i)di\right)d\iota \qquad (9) \end{aligned}$$

where p^1 is the first-order derivative of the antagonism function with respect to identification.

Let $z_\iota \geq 0$ be an identification threshold and let $H_A(z_\iota) = \int_{z_\iota}^{\infty} f_A(i)di$ denote the identification dominance curve function in a population A. $H_A(z_\iota)$ is the proportion of individuals for whom identification in their groups is greater than z_ι. We have $\int_0^{\infty} f(i)di = H_A(0) = 1$ and

$$P_A = p(0) + \int_0^{\infty} p^1(\iota)H_A(\iota)d\iota. \qquad (10)$$

The identification dominance curve function $H(z_\iota)$ is defined analogously for a distribution B:

$$P_B \quad = \quad p(0)+\int_0^\infty p^1(\iota)H_B(\iota)d\iota. \tag{11}$$

Let E^1 denote the class of polarization indices in the form of Equation (5) where the antagonism function is increasing in identification.

$$E^1=\{P \text{ in the form of Equation (5)} \mid p^1\geq 0\}. \tag{12}$$

We are searching conditions under which distribution A is more homogenous (always has more identification) than distribution B regardless of the polarization index in E^1, that is, conditions under which $P_A-P_B\geq 0$ for all $P\in E^1$:

$$P_A-P_B=\int_0^\infty p^1(\iota)\left(H_A(\iota)-H_B(\iota)\right)d\iota. \tag{13}$$

Consider $P \in E^1$ and suppose that $H_A(z_\iota)-H_B(z_\iota)\geq 0$ whatever the identification threshold $z_\iota\geq 0$. Since $p^1\geq 0$, $\Delta P=P_A-P_B\geq 0$ as an integral of a positive function in a closed interval. That is, the condition $H_A(z_\iota)-H_B(z_\iota)\geq 0$ for all $z_\iota\geq 0$ is a sufficient condition to compare distributions A and B with polarization indices in E^1.

For the necessity condition, suppose that $P_A-P_B\geq 0$ for all $P\in E^1$ and that on the other side, $H_A(z_\iota)-H_B(z_\iota)<0$ for $z_\iota\in[\bar{z}_\iota,\bar{z}_\iota+\varepsilon]$ for a small value of $\varepsilon>0$. We define an antagonism function by:

$$p(\iota)=\begin{cases}\bar{z}_\iota & \text{if } \iota\leq\bar{z}_\iota\\ \iota & \text{if } \bar{z}_\iota<\iota\leq\bar{z}_\iota+\varepsilon\\ \bar{z}_\iota+\varepsilon & \text{if } \iota>\bar{z}_\iota+\varepsilon\end{cases}$$

The first-order derivative of the antagonism function with respect to identification is $p^1(\iota)=\frac{\partial p(\iota)}{\partial\iota}=1$ in $[\bar{z}_\iota,\bar{z}_\iota+\varepsilon]$, and 0 elsewhere. So, equation Equation (13) becomes:

$$P_A-P_B=\int_{\bar{z}_\iota}^{\bar{z}_\iota+\varepsilon}\left(H_A(\iota)-H_B(\iota)\right)d\iota. \tag{14}$$

With the hypothesis that $H_A(z_\iota)-H_B(z_\iota)<0$ in $[\bar{z}_\iota,\bar{z}_\iota+\varepsilon]$, we also have $\int_{\bar{z}_\iota}^{\bar{z}_\iota+\varepsilon}\left(H_A(\iota)-H_B(\iota)\right)d\iota<0$ which implies that $P_A-P_B<0$ and contradicts the assumption of $P_A-P_B\geq 0$. The necessary condition is then verified.

Distribution B is therefore said to stochastically dominate distribution A at first order in the identification component of polarization if $H_A(z_\iota)-H_B(z_\iota)\geq 0$ for all $z_\iota\geq 0$. When the above condition is fulfilled, there are always more groups with densities greater than any identification threshold in distribution A than in distribution B, and the identification dominance curve of A is always above that of B.

As a result, income groups in distribution A have more similar individuals and are more homogenous than income groups in distribution B. We note $B \succcurlyeq A$ to mean that B is less homogenous than A and stochastically dominates A at first order.

Proposition 3.1 . E^1 ***identification dominance at first order***: *Given two distributions from two distinct populations A and B,*
$B \succcurlyeq A$ *for all* $P \in E^1$ *iff* $H_A(z_\iota) - H_B(z_\iota) \geq 0$ *for all* $z_\iota \geq 0$.

3.2 Second-order stochastic dominance

A distribution B is less homogenous and stochastically dominates another distribution A at first order if the identification dominance curve of B, $H_B(z_\iota)$, always lies below the identification dominance curve of A, $H_A(z_\iota)$, across a range of identification thresholds z_ι. However, instead of having dominance in the entire interval of identification thresholds, distribution B might stochastically dominate distribution A at first order only from a certain value of identification threshold z_ι^1.

Two situations can occur when distribution B stochastically dominates distribution A: Either dominance holds in the entire interval of identification thresholds, or else dominance reverses at the value z_ι^1 where the two identification dominance curves intersect (Figure 2). The first case corresponds to the dominance in the sense of Foster and Shorrocks (1988) while the second case implies the definition of z_ι^1 as

$$z_\iota^1 \quad = \quad sup\{z_\iota > 0 | H_A(z_\iota) < H_B(z_\iota)\}. \tag{15}$$

The situation where $z_\iota^1 > 0$ opens a possibility for second-order stochastic dominance. To see how, consider the expression of the polarization index from Equation (10):

$$P_A \quad = \quad p(0) + \int_0^\infty p^1(i) H_A(i) di. \tag{16}$$

Setting $\iota = -i$ yields:

$$P_A \quad = \quad p(0) + \int_{-\infty}^0 p^1(-\iota) H_A(-\iota) d\iota. \tag{17}$$

Then, integrating by parts leads to:

$$\begin{aligned} P_A \quad &= \quad p(0) + p^1(\iota) \int_{-\infty}^{\iota} H_A(-i) di|_{-\infty}^{0} + \int_{-\infty}^{0} p^2(-\iota) \left(\int_{-\infty}^{\iota} H_A(-i) di \right) d\iota & (18) \\ &= \quad p(0) + p^1(0) \int_{-\infty}^{0} H_A(-i) di + \int_{-\infty}^{0} p^2(-\iota) \left(\int_{-\infty}^{\iota} H_A(-i) di \right) d\iota & (19) \\ &= \quad p(0) + p^1(0) \int_{0}^{\infty} H_A(i) di + \int_{0}^{\infty} p^2(\iota) \left(\int_{\iota}^{\infty} H_A(i) di \right) d\iota & (20) \end{aligned}$$

where p^2 is the second-order derivative of the antagonism function with respect to identification.

Let $H^2(\iota) = \int_\iota^\infty H(i)di$ be defined on the support of identification. Since $H^2(0) = 1$, the above development of the polarization index becomes:

$$\begin{aligned} P_A &= p(0) + p^1(0)H_A^2(0) + \int_0^\infty p^2(\iota)H_A^2(\iota)d\iota \\ &= p(0) + p^1(0) + \int_0^\infty p^2(\iota)H_A^2(\iota)d\iota. \end{aligned} \tag{21}$$

The analogous quantity can be written for a distribution B:

$$P_B = p(0) + p^1(0) + \int_0^\infty p^2(\iota)H_B^2(\iota)d\iota.$$

$$\text{Thus, } \Delta P = \int_0^\infty p^2(\iota)(H_A^2(\iota) - H_B^2(\iota))d\iota. \tag{22}$$

Let E^2 denote the class of polarization indices in the form of Equation (5) where the antagonism function is convex in identification.

$$E^2 = \{P \text{ in the form of Equation (5)} | \; p^2 \geq 0\}. \tag{23}$$

We are searching conditions under which $\Delta P = P_A - P_B \geq 0$ for all $P \in E^2$. Consider $P \in E^2$ and suppose that $H_A^2(z_\iota) - H_B^2(z_\iota) \geq 0$ regardless of the identification threshold $z_\iota \geq 0$. Since $p^2 \geq 0$, $\Delta P \geq 0$ for the same reasons as above. That is, the condition $H_A^2(z_\iota) - H_B^2(z_\iota) \geq 0$ for all $z_\iota \geq 0$ is a sufficient condition for second-order stochastic dominance.

For the necessity condition, suppose that $P_A - P_B \geq 0$ for all $P \in E^2$ and that $H_A^2(z_\iota) - H_B^2(z_\iota) < 0$ for $z_\iota \in [\bar{z}_\iota, \bar{z}_\iota + \varepsilon]$ for a small value of $\varepsilon > 0$. We define an antagonism function by: $p(\iota) = \begin{cases} \bar{z}_\iota^2 & \text{if } \iota \leq \bar{z}_\iota \\ \iota^2 & \text{if } \bar{z}_\iota < \iota \leq \bar{z}_\iota + \varepsilon \\ (\bar{z}_\iota + \varepsilon)^2 & \text{if } \iota > \bar{z}_\iota + \varepsilon \end{cases}$

The second-order derivative of the antagonism function with respect to identification is $p^2(\iota) = 2$ in $[\bar{z}_\iota, \bar{z}_\iota + \varepsilon]$, and 0 elsewhere. So, equation Equation (22) becomes:

$$P_A - P_B = \int_{\bar{z}_\iota}^{\bar{z}_\iota + \varepsilon} 2\left(H_A^2(\iota) - H_B^2(\iota)\right) d\iota. \tag{24}$$

With the hypothesis that $H_A^2(z_\iota) - H_B^2(z_\iota) < 0$ in $[\bar{z}_\iota, \bar{z}_\iota + \varepsilon]$, we have $\int_{\bar{z}_\iota}^{\bar{z}_\iota + \varepsilon} \left(H_A^2(\iota) - H_B^2(\iota)\right) d\iota < 0$ which implies that $P_A - P_B < 0$ and contradicts the assumption of $P_A - P_B \geq 0$. The necessary condition is then verified. We note $B \succcurlyeq_2 A$ to mean that B stochastically dominates A at second order.

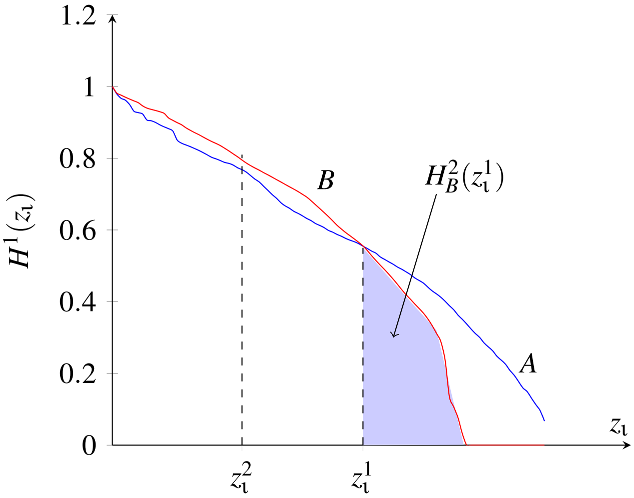
Proposition 3.2 . E^2 ***identification dominance at second order****: Given two distributions from two distinct populations A and B,*
$B \succcurlyeq_2 A$ for all $P \in E^2$ iff $H_A^2(z_\iota) - H_B^2(z_\iota) \geq 0$ for all $z_\iota \geq 0$.

As in first-order stochastic dominance, two situations can occur: either B stochastically dominates A at second order everywhere, or else there exits z_ι^2 (see Figure 2) defined by

$$z_\iota^2 \quad = \quad sup\{z_\iota > 0 | H_A^2(z_\iota) < H_B^2(z_\iota)\} \tag{25}$$

from which B stochastically dominates A at second order. This suggests to check for dominance at order three.

Figure 2: Identification dominance curves for two distributions A and B



3.3 Higher-order stochastic dominance

The above procedure can be developed for stochastic dominance at order $s > 2$. The polarization index integrated by part s times with respect to identification leads to :

$$P \quad = \quad p(0) + p^1(0) + \cdots + p^{s-1}(0) \int_0^\infty p^s(\iota) H^s(\iota) d\iota.$$

where p^s is the s^{th}-order derivative of the antagonism function with respect to identification, $H^s(\iota) = \int_\iota^\infty H^{s-1}(i) di$ for $s \geq 2$ and $H^1(\iota) = H(\iota)$.

For two populations A and B,

$$\Delta P \quad = \quad \int_0^\infty p^s(\iota)(H_A^s(\iota) - H_B^s(\iota)) d\iota. \tag{26}$$

Let E^s denote the sub class of polarization indices in the form of Equation (5) defined as:

$$E^s = \{P \text{ in the form of Equation (5) } | p^s \geq 0\}. \tag{27}$$

Suppose that $H_A^s(z_\iota) - H_B^s(z_\iota) \geq 0$ for all $z_\iota \geq 0$, then $\Delta P \geq 0$ which fulfills the sufficient condition. For the necessity condition, assume $P_A - P_B \geq 0$ and consider the following function of antagonism.

$$p(\iota) = \begin{cases} \bar{z}_\iota^s & \text{if } \iota \leq \bar{z}_\iota \\ \iota^s & \text{if } \bar{z}_\iota < \iota \leq \bar{z}_\iota + \varepsilon \\ (\bar{z}_\iota + \varepsilon)^s & \text{if } \iota > \bar{z}_\iota + \varepsilon \end{cases}$$

Its s^{th}-order derivative with respect to identification is $p^s(\iota) = s!$ in $[\bar{z}_\iota, \bar{z}_\iota + \varepsilon]$, and 0 elsewhere. Then:

$$P_A - P_B = \int_{\bar{z}_\iota}^{\bar{z}_\iota + \varepsilon} s!\,(H_A^s(\iota) - H_B^s(\iota))\,d\iota. \tag{28}$$

Suppose $H_A^s(z_\iota) - H_B^s(z_\iota) < 0$, then $\int_{\bar{z}_\iota}^{\bar{z}_\iota+\varepsilon} (H_A^s(\iota) - H_B^s(\iota))\,d\iota < 0$, which implies that $P_A - P_B < 0$ and contradicts the assumption of $P_A - P_B \geq 0$. The necessary condition is then verified. We note $B \succcurlyeq_s A$ to mean that B stochastically dominates A at s^{th} order.

Proposition 3.3 . E^s ***identification dominance at order*** **s**: *Given two distributions from two distinct populations A and B,*
$B \succcurlyeq_s A$ for all $P \in E^s$ iff $H_A^s(z_\iota) - H_B^s(z_\iota) \geq 0$ for all $z_\iota \geq 0$.

For each dominance order s, let an identification limit z_ι^s defined by

$$z_\iota^s = sup\{z_\iota > 0 | H_A^s(z_\iota) < H_B^s(z_\iota)\}. \tag{29}$$

The procedure stops either when the value z_ι^s is judged to be sufficiently high or when we reach the maximum value of z_ι. In the first case, z_ι^s has become greater that a reasonable maximum value of identification threshold. In the second case, stochastic dominance is achieved everywhere.

4 Case of income distributions

In this section, the identification dominance curve $H(z_\iota)$ is explained in the space of income where identification is measured with a density of income:

$$\iota(y) = f(y). \tag{30}$$

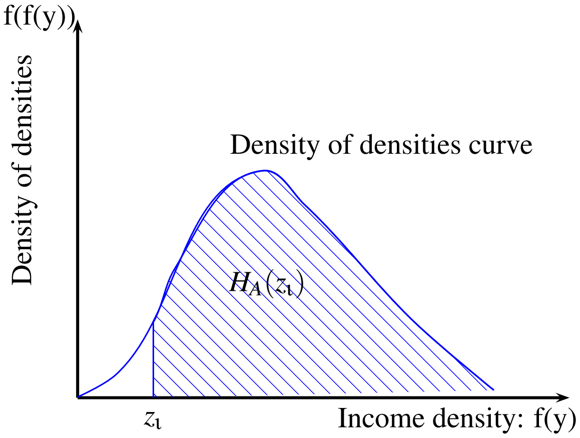
$H(z_\iota)$ becomes:

$$H(z_\iota) = \int_{z_\iota}^{\infty} dF_f(\chi) \tag{31}$$

$$= 1 - F_f(z_\iota) \tag{32}$$

where F_f is the cumulative distribution function of income densities and $H(z_\iota)$ is the hatched area of Figure 3.

Figure 3: Identification dominance curve for z_ι, $H(z_\iota)$: Proportion of individuals whose income density (identification) is greater than z_ι



5 Estimation and inference

This section is devoted to statistical inference of dominance relations between two identification dominance curves. It arrives that dominance observed in two distributions using dominance curves is induced by noise in samples. Furthermore, when two curves are close, we cannot conclude about the presence of dominance even if one is always above the other. We propose an estimator of $H(z_\iota)$ and we test dominance using an Intersection-Union [IU] test. The IU test introduced by Berger (1982) and used by Berger and Sinclair (1984) and Berger (1996) consists of choosing the null hypothesis as a union of simple hypotheses and the alternative hypothesis as an intersection of simple hypotheses.

5.1 Distribution and critical frontiers

Consider a random sample of N independently and identically distributed observations of income $y_1, \cdots, y_N$ drawn from a distribution with a cumulative distribution function F. An estimator of $H(z_\iota)$ is defined by :

$$\begin{aligned} \hat{H}(z_\iota) &= \int I(\hat{f}(y) \geq z_\iota) d\hat{F}_f(\chi) \\ &= N^{-1} \sum_{i=1}^{N} I(\hat{f}(y_i) \geq z_\iota). \end{aligned} \tag{33}$$

The income density f is estimated non parametrically with the Gaussian kernel function:

$$\hat{f}(y_i) \quad = \quad \frac{N^{-1}}{h\sqrt{2\pi}} \sum_{j=1}^{N} exp^{-0.5\left(\frac{y_j - y_i}{h}\right)^2}$$

where h is the optimal bandwidth of Silverman, proposed to trade off reductions of bias and variance:

$$h \quad = \quad 1.06\hat{\sigma}N^{-\frac{1}{5}}$$

where $\hat{\sigma}$ is an estimator of the standard deviation of income, which is conditional on the gap between the first and the third quartile. Let $Q(0.25), Q(0.75)$ and σ denote respectively the estimates of the first and the third quartiles, and the standard deviation of the distribution. Then,

$$\hat{\sigma} = \begin{cases} \frac{Q(0.75)-Q(0.25)}{1.34} & \text{if } \frac{Q(0.75)-Q(0.25)}{1.34} < \sigma \\ \sigma & \text{if } \frac{Q(0.75)-Q(0.25)}{1.34} \geq \sigma. \end{cases} \tag{34}$$

The existence of the population moment of order 1 allows us to apply the law of large numbers to Equation (33) and deduce that $\hat{H}(z_\iota)$ is a consistent estimator of $H(z_\iota)$. The central limit theorem allows us to assert that this estimator is root-N consistent and asymptotically normal. Finally, the asymptotic distribution of the estimator of the identification dominance curve is defined in Theorem 5.1:

Theorem 5.1 *$N^{\frac{1}{2}}(\hat{H}(z_\iota) - H(z_\iota))$ is asymptotically normal with mean zero, and the structure of the variance is given by:*

$$\begin{aligned} \lim_{N\to\infty} NVar(\hat{H}(z_\iota) - H(z_\iota)) \quad &= \quad (1 - F_f(z_\iota))F_f(z_\iota) + J^{-2}\sum_{j=1}^{J} Var(\theta_j(z_\iota)) \\ &+ \quad 2J^{-1}(1 - F_f(z_\iota))\sum_{j=1}^{J}[E(\theta_j(z_\iota)|f(y) \geq z_\iota) - E(\theta_j(z_\iota))] \end{aligned} \tag{35}$$

where there exist $y_1^*, \cdots, y_j^*, \cdots, y_J^* | f(y_j^*) = z_\iota$,

$$\theta_j(z_\iota) \quad = \quad \hat{f}(y_j^*)f_{f,j}(z_\iota), \tag{36}$$

$$f_{f,j}(z_\iota) \quad = \quad N^{-1}\sum_{i=1}^{N} \frac{k_p(\frac{y_i - y_j^*}{h})}{J^{-1}\sum_{j=1}^{J} k_p(\frac{y_i - y_j^*}{h})} k_f\left(\frac{f(y_i) - z_\iota}{h}\right) \tag{37}$$

where k_p and k_f are Gaussian kernel functions defined respectively in the income and density spaces,

$$Var(\theta_j(z_\iota)) \quad = \quad f_{f,j}^2(z_\iota)\frac{N^{-1}}{h} f(y_j^*) \int k_p^2(\psi_j)d\psi_j \quad \text{where } \psi_j = \frac{y - y_j^*}{h}, \tag{38}$$

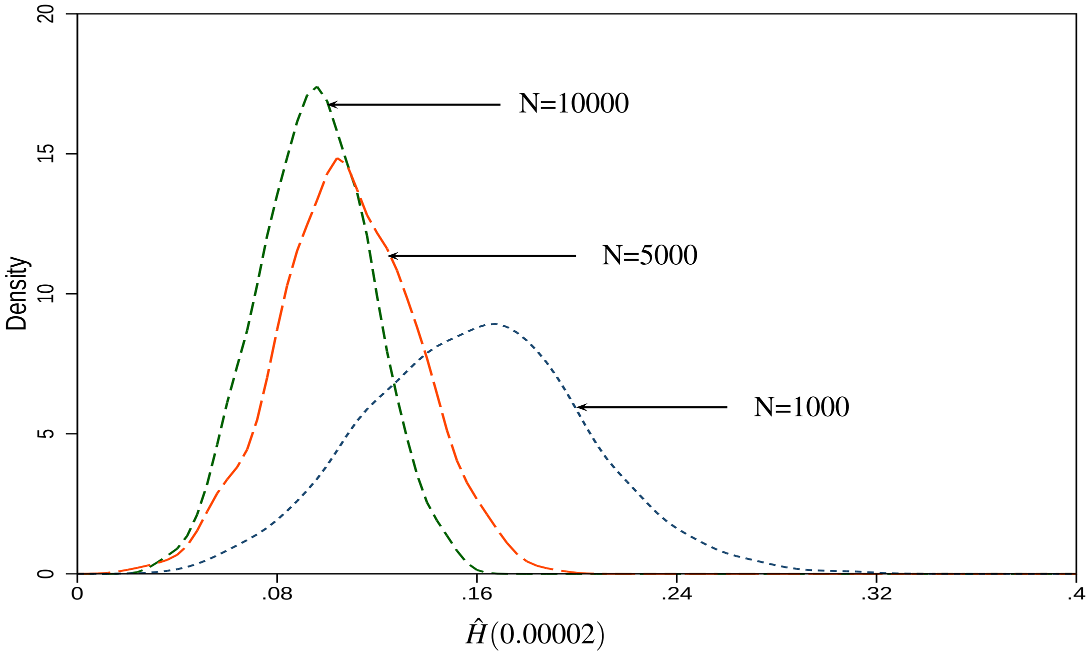
$$E(\theta_j(z_\iota)|f(y) \geq z_\iota) \quad = \quad f_{f,j}(z_\iota)E(\hat{f}(y_j^*)|f(y) \geq z_\iota), \tag{39}$$

$$\text{and} \quad E(\theta_j(z_\iota)) \quad = \quad f_{f,j}(z_\iota)\int k_p(\psi_j)f(h\psi_j + y_j^*)d\psi_j. \tag{40}$$

(Proofs are given in the appendix).

We also study the convergence in probability of the estimator $\hat{H}(z_\iota)$ through a Monte Carlo simulation. We first take an identification threshold $z_\iota = 0.00002$. Next we draw three sets of 1000 samples from a log-normal distribution of mean zero and standard deviation 1 with sample sizes set respectively to 1000, 5000 and 10 000 observations. We draw the density of the estimators obtained from the 1000 simulated samples. In Figure 4, the density curves narrow in as the sample size increases. This tendency shows that the estimator $\hat{H}(z_\iota)$ is convergent. The distribution of the estimator is biased on the right when the sample size is 1000. We attribute this tendency to the fact the gaussian kernel estimator is higher and less concentrated in samples with small sizes. The gaussian kernel estimator is not efficient when the size of a sample is small.

Figure 4: Density curves showing the convergence of $\hat{H}(z_\iota)$ when $z_\iota = 0.00002$. For each sample size N, 1000 samples are drawn from a log-normal distribution of mean zero and standard deviation 1



The identification dominance curve $H(z_\iota)$ equals 1 when $z_\iota = 0$ and 0 for large values of z_ι in any distribution. Thus, at the lower and upper bounds of identification thresholds, two identification dominance curves are confounded and dominance is expected in a restricted interval of identification thresholds. The appropriate range of thresholds that can order distributions can be estimated. Consider two income distributions from two populations A and B where the interest is about the dominance of A by B. For distribution B to dominate distribution A, there must exist a threshold $z_\iota^- > 0$ from which $H_A(z_\iota)$ begins to be greater than $H_B(z_\iota)$:

$$z_\iota^- \quad = \quad sup\{z_\iota > 0 | H_A(z_\iota) < H_B(z_\iota)\}. \tag{41}$$

From this lower critical frontier z_ι^-, either $H_B(z_\iota) \leq H_A(z_\iota)$ for all $z_\iota \geq z_\iota^-$ and distribution B stochastically dominates distribution A everywhere above z_ι^-, or else there exists a reversal identification threshold $z_\iota^+ > z_\iota^-$ up to which dominance applies. The upper critical frontier z_ι^+ is the maximum identification threshold up to which distribution B stochastically dominates distribution A at first order. It is defined by:

$$z_\iota^+ \quad = \quad sup\{z_\iota > z_\iota^- | H_B(z_\iota) \leq H_A(z_\iota)\}. \tag{42}$$

5.2 Testing dominance

Testing for dominance implies addressing two issues, that of the choice of the null and alternative hypotheses and that of the test statistic. Davidson and Duclos (2000) present three nested hypotheses that could be used either as the null or as the alternative hypothesis in dominance tests. The first one is $H_0 : H_A(z_\iota) - H_B(z_\iota) = 0 \;$ for all $z_\iota \leq z_\iota^+$. The second one is $H_1 : H_A(z_\iota) - H_B(z_\iota) \geq 0 \;$ for all $z_\iota \leq z_\iota^+$ and the third and last one is H_2 without any restriction on $H_A(z_\iota) - H_B(z_\iota)$. The three hypotheses are nested in the sense that $H_0 \subset H_1 \subset H_2$.

McFadden (1989) proposes to test H_0 against H_1 with the test statistic of the supremum. It is a form of the Kolmogorov-Smirnov test used to test the identity of two distributions. The main difficulty in the test of McFadden is to tract the asymptotic properties of the test statistic under the null hypothesis. Also, it supposes that the two samples under consideration have the same sizes, which is not always the case.

The approach proposed by Kaur et al. (1994) [KPS] seems more appropriate. The minimum value of test statistics at different alienation thresholds is used as the test statistic for the null of nondominance against the alternative of dominance. One advantage of this formulation is that if we reject the null hypothesis, all that remains is the alternative of dominance (Davidson and Duclos, 2013). This test is the Intersection-Union test and the probability of rejection of the null when it is true is shown to be asymptotically bounded by the nominal level of a test based on the standard normal distribution (Davidson and Duclos, 2000).

So, we propose the following test of H_0^ι against H_1^ι to test the dominance of A by B:

$$\begin{cases} H_0^\iota : H_B(z_\iota) \geq H_A(z_\iota) \text{ for some } z_\iota \in [z_\iota^-,\; z_\iota^+] \\ H_1^\iota : H_B(z_\iota) < H_A(z_\iota) \text{ for all } z_\iota \in [z_\iota^-,\; z_\iota^+]. \end{cases} \tag{43}$$

A procedure to perform this test consists of using a predetermined grid of points corresponding in our case to identification thresholds (Howes, 1993). The only precaution is to be sure to cover entirely the interval of interest. The properties of the IU test using a grid of points are similar to those of the KPS test.

The null hypothesis is a union of simple hypotheses and the alternative hypothesis is an intersection of simple hypotheses. In this case, evidence against the null hypothesis exists when it also exists for every simple null hypothesis. If each simple test is of level α, then the IU test is also of level α following theorems 1 and 2 of Berger (1982). Furthermore, Berger and Sinclair (1984) show that not only the IU test is of level α, but it is exactly of size α.

We suppose that the two distributions are independent. The KPS statistic is defined as the minimum over alienation thresholds of the simple test statistics. Consider the following simple test (of $H_0^{z_\iota}$ against $H_1^{z_\iota}$) for the dominance of A by B when the identification threshold is z_ι:

$$\begin{cases} H_0^{z_\iota} : H_B(z_a) - H_A(z_\iota) \geq 0 \\ H_1^{z_\iota} : H_B(z_\iota) - H_A(z_\iota) < 0 \end{cases} \tag{44}$$

The test statistic associated to this simple test is

$$t_{z_\iota} = \frac{\hat{H}_B(z_\iota) - \hat{H}_A(z_\iota)}{\sqrt{Var(\hat{H}_A(z_\iota)) + Var(\hat{H}_A(z_\iota))}}$$

Because of the assumption on the independence of the two distributions, the variance of $\hat{H}_A(z_\iota) - \hat{H}_B(z_\iota)$ is the sum of their variances. Under the null hypothesis, there is no loss to take $H_A(z_\iota) = H_B(z_\iota)$, a condition under which t_{z_ι} tends asymptotically to the standard normal distribution $\mathcal{N}(0,1)$. At a nominal level α, there is strong evidence against the null hypothesis of this simple test if the test statistic t_{z_ι} is lower than the critical value of the standard normal distribution at the level α. The test statistic of the IU test is defined by: $t = min\{t_{z_\iota}\}$ for z_ι in the grid of thresholds . There is strong evidence against the null hypothesis of the IU test at the nominal level α if there is strong evidence against all simple tests. In that case, the test statistic t is lower than the critical value of the standard normal distribution at the level α.

The evidence against the null hypothesis can be tested using p-values. A p-value is defined as a measure of the strength of the evidence against the null hypothesis and is determined as the probability of getting a value greater than or equal to the test statistic. We perform statistical test using a bootstrap approach because it gives better results than asymptotic tests, being more efficient. This choice avoids facing the problem of the bias in the estimator for samples of small sizes:

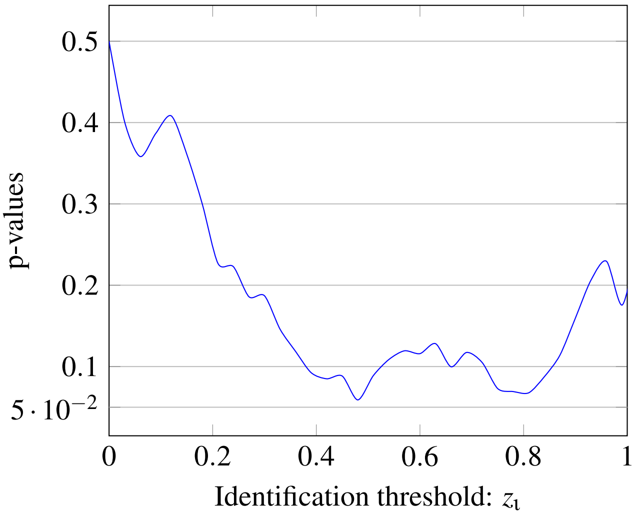
consider two income distributions A and B where we want to test the dominance of A by B, and a grid $z_\iota^1, z_\iota^2, \cdots, z_\iota^k, \cdots, z_\iota^M$ of M points of identification thresholds. Here is the procedure to compute bootstrapped p-values.

1. consider an identification threshold z_ι^k on the grid, say the first one;
2. draw a pair of samples with replacement from A and B of the same sizes as the initial samples and compute $\hat{d}_1^k = \hat{H}_B(z_\iota^k) - \hat{H}_A(z_\iota^k)$;
3. repeat the previous step (step 2) S independent times;
4. for each sample pair $s = 1, 2, \cdots, S$, keep the estimate $\hat{d}_s^k = \hat{H}_B(z_\iota^k) - \hat{H}_A(z_\iota^k)$;
5. the p-value of the simple test with z_ι^k is given by $p_k = S^{-1}\sum_{s=1}^{S} I(\hat{d}_s^k \geq 0)$ (Duclos and Araar, 2006);
6. for the rest of identification thresholds $z_\iota^k, k = 2, \cdots, M$, repeat steps 2 to 5.

The result is a pair (z_ι^k, p_k) of identification thresholds and corresponding p-values. The p-value p_k indicates the maximum probability that an error is made when we reject the null hypothesis in favour of the alternative in the simple test with z_ι^k. For a significance level α, the null hypothesis of the simple test is rejected if $p_k \leq \alpha$.

A decision on the IU test in Equation 43 can be taken graphically by drawing the curve of p-values as a function of identification thresholds. Dominance exists where the p-value curve lies beneath the level of the test. For instance, consider Figure 5 below of the dominance test between two hypothetical distributions A and B, on which we have p-values on the vertical axis and identification thresholds on the horizontal axis. For this illustrative example and for the traditional significance level of 5%, the null hypothesis is never rejected since p-values are always greater than 0.05 regardless of the threshold. However, the null hypothesis is sometimes rejected when the significance level is set to 10%. In this last case, the null hypothesis cannot be rejected until the threshold reaches 0.39. From this threshold, p-values remain less than 10% until the threshold is 0.51. The critical frontiers can then be fixed respectively to $z_\iota^- = 0.39$ and $z_\iota^+ = 0.51$, values between which distribution B stochastically dominates distribution A.

Figure 5: Graphical illustration of a statistical test with p-value at different identification thresholds



6 Empirical illustration using LIS data

6.1 Data and statistics

We apply the above methodology to 2004 data of 11 countries from the Luxembourg Income Study (LIS) database (see Table 1). The LIS database contains harmonized microdata from countries which are exclusively middle and high-income countries. It also contains variables on household income, taxes and transfers, households and individual characteristics, labor market outcomes, and expenditures. We use disposable household income, defined as post tax and post transfer income of the household. The household income is adjusted through the division by an adult equivalent scale $s^{0.5}$ where s is the household size. This operation is introduced to provide comparable incomes for individuals living in households of different sizes. We also weighted all household observations by the household weight "hweight normalized" multiplied by the number of people in the household unit. Some (less than 0.5%) incomes are negative in each country. We could have set these values to zero but we did not, assuming their impact is not significant on results. Income is also normalized by the mean to eliminate the problem of measurement units.

To compute some statistics like the mean and median incomes, national currencies are converted into US $ with 2005 purchasing power parities (PPP) from the 2005 International Comparison Program. Sample sizes for each country are shown in the first column of Table 1. This table also summarizes the mean and the median of the adjusted household net income of each country set in US $. The USA has the highest mean income ($31,616.15) among the 11 countries. It is followed by Switzerland with a mean of $29,205.55. The lowest mean income is that of Peru ($4,761.52). For countries like Mexico and Poland, the mean incomes are less than 10,000 dollars, but are greater than that of Peru. The Switzerland median income found to be $26,652 is the highest followed by the USA with a median income found to be $25,696.50. The lowest values of the median income are those of Peru ($2,964.09) and Mexico ($4,425.04).

Table 1: Sample sizes, mean and median income for 11 countries. Year=2004.

Sigle	Country	Sample size	$US, ppp=2005	
			Mean	**Median**
ca	Canada	27,820	26,886.37	23,103.57
dk	Denmark	83,349	22,197.02	20,504.93
fi	Finland	11,229	20,784.08	18,199.34
ie	Ireland	6,080	23,171.57	19,431.43
mx	Mexico	22,595	6,938.19	4,425.04
no	Norway	13,131	27,754.11	24,599.64
pe	Peru	18,383	4,761.52	2,964.09
pl	Poland	32,214	7,975.03	6,837.35
si	Slovenia	3,725	15,494.17	14,433.74
ch	Switzerland	3,270	29,205.17	26,652.3
us	United States	76,447	31,616.15	25,696.5

The 2005 PPP$_s$ come from the 2005 International Comparison Program

6.2 Results

The application is limited to stochastic dominance at first order. The results are presented in two parts. The first part presents comparisons of pairs between the 11 countries based on identification dominance curves. A country whose identification dominance curve is always above another one regardless of the identification threshold has more identification and is more homogenous. A country B stochastically dominates at first order another country A when there exist z_ι^- and z_ι^+ for which $H_B(z_\iota) \leq H_A(z_\iota)$ for all $z_\iota \in [z_\iota^-, z_\iota^+]$. Then income groups in B are less homogenous than income groups in A, or individuals belonging to a same income group in A feel more alike than individuals belonging to a same income group in B.

Using identification dominance curves like those in Figure 7, we see that the USA stochastically dominates (is less homogeneous than) all the other countries except Mexico and Peru (Figure 6). Apart from these two countries, $H_{USA}(z_\iota) \leq H_{Country}(z_\iota)$ for any identification threshold z_ι. The proportion of individuals whose income densities are greater than any identification threshold is lower in the USA than in the other countries. That is, people in Canada, Denmark, Finland, Ireland, Norway, Poland, Slovenia and Switzerland feel more alike and are more homogenous than people in the USA. As a result, income distributions in these countries are more homogenous than the income distribution in the USA. The results for pairwise comparisons are presented in Table 2.

For small values of identification thresholds, the USA is more homogeneous than Mexico and Peru, but the situation reverses for higher values of identification thresholds (Figure 6). Mexico stochastically dominates at first order the USA for $z_\iota \in]0,0.6]$ and Peru stochastically dominates at first order the USA for $z_\iota \in]0,0.77]$. The identification dominance curves of the USA drawn across these intervals are always above those of Mexico and Peru, and outside these intervals, the identification dominance curves of the USA are always below those of the two countries.

Canada stochastically dominates everywhere at first order Denmark, Finland, Ireland, Norway, Slovenia and Switzerland. Then, the income distribution in Canada is less homogeneity than income distributions in these countries. As a consequence, the identification dominance curve of Canada is always below those of these countries. However, Canada is stochastically dominated by Mexico and Peru in restricted intervals. Mexico stochastically dominates (is less homogeneous than) Canada at first order for $z_\iota \in [0,0.74]$. So, in Canada, individuals whose income densities are lower than or equal to 0.74 feel more alike than the corresponding individuals in Mexico. Likewise Peru stochastically dominates Canada at firs order for $z_\iota \in [0,0.85]$. Further, Canada is as homogeneous as Poland.

Denmark is stochastically dominated at fist order by all the other countries, that it is more homogeneous than the rest of the countries. Dominance of Denmark is achieved everywhere by Ireland, Mexico, Peru, Poland, Switzerland and the USA. Income distributions in these countries are less homogeneous than the income distribution in Denmark. Then the income distribution in Denmark leads to groups of individuals who feel more alike compared to income distributions in Ireland, Mexico, Peru, Poland, Switzerland and the USA. Countries as Finland, Norway and Slovenia dominate Denmark in restricted intervals. Dominance is achieved in [0, 0.93] for Finland, [0, 0.92] for Norway and [0, 0.94] for Slovenia. For higher densities of income, Denmark appeared to be less homogeneous than the above countries. In fact, curves are confounded in the queues of distributions and the decision about dominance is ambiguous due to the small number of observations in the extreme of distributions.

The income distribution in Finland is as homogenous as the income distribution in Slovenia, meaning that individuals belonging to a same income group in Finland feel as alike as those from the same income group in Slovenia. Finland is stochastically dominated everywhere by Ireland, Mexico, Peru, Poland and Switzerland at first order, but Finland stochastically dominates Norway everywhere at first order. In Norway the income distribution is more homogenous than the income distribution in Finland where the income in turn is more homogeneous than those of the other countries.

Ireland is less homogeneous than Norway and Slovenia, but it is more homogeneous than Mexico, Peru, Poland, Switzerland and the USA. In fact, Ireland stochastically dominates everywhere Norway and Slovenia, but it is stochastically dominated by Mexico, Peru, Poland, Switzerland and the USA. Dominance by Switzerland and the USA is achieved everywhere while dominance by Mexico, Peru and Poland is achieved in restricted intervals of identification thresholds. The identification thresholds up to which these last three countries dominate Ireland are found to be respectively 0.76, 0.79 0.72. Above these values, they are dominated by Ireland. Mexico is less homogeneous than Norway, Poland, Slovenia and Switzerland and is more homogeneous than Peru.

Norway is less homogeneous than Slovenia and is more homogeneous than Peru, Poland, Switzerland and the USA. Peru is less homogeneous than Poland, Slovenia, Switzerland and the USA, Poland is less homogeneous than Slovenia, Switzerland and the USA, and finally Slovenia is more homogeneous than Switzerland and the USA.

Table 2: Result of identification dominance between 11 countries based on their identification dominance curves

	ca	dk	fi	ie	mx	no	pe	pl	si	ch
dk	$\preccurlyeq$									
fi	$\preccurlyeq$	${}_r\succcurlyeq$								
ie	$\preccurlyeq_r$	$\succcurlyeq$	$\succcurlyeq$							
mx	${}_r\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$	${}_r\succcurlyeq$						
no	$\preccurlyeq$	${}_r\succcurlyeq$	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$					
pe	${}_r\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$	${}_r\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$				
pl	$\approxeq$	$\succcurlyeq$	$\succcurlyeq$	${}_r\succcurlyeq$	$\preccurlyeq$	$\succcurlyeq$	$\preccurlyeq$			
si	$\preccurlyeq$	${}_r\succcurlyeq$	$\approxeq$	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$		
ch	$\preccurlyeq$	$\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$	$\preccurlyeq$	$\succcurlyeq$	$\preccurlyeq$	$\preccurlyeq$	$\succcurlyeq$	
us	$\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$	$\preccurlyeq_r$	$\succcurlyeq$	$\preccurlyeq_r$	$\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$

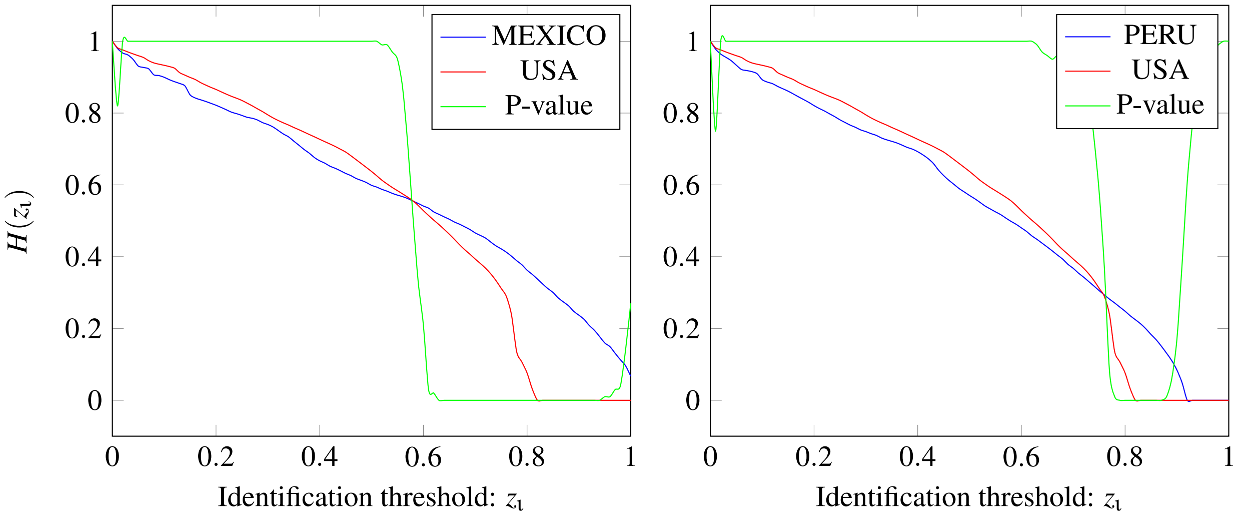
r $\equiv$ *dominance is not achieved everywhere, but in a restricted interval*

In a cell, the sign $\succcurlyeq$ means that the country in line stochastically dominates at first order the country in column; the sign $\preccurlyeq$ means that the country in line is stochastically dominated at first order by the country in column; the sign $\approxeq$ means that the curve of the country in line is close (in the manner of the curves of Mexico and the USA before the crossover point) to the one of the country in column.

The second part of the results presents dominance tests between the USA and the rest of 10 countries. Results are drawn graphically using bootstrapped p-values. The number of replication in bootstrap tests is $B = 100$ for any identification threshold. We test if the USA is less homogeneous and stochastically dominates another country at a level of 5%. The null hypothesis is then the non-dominance of a country by the USA. When the USA dominates, p-values are lower than the level of the test.

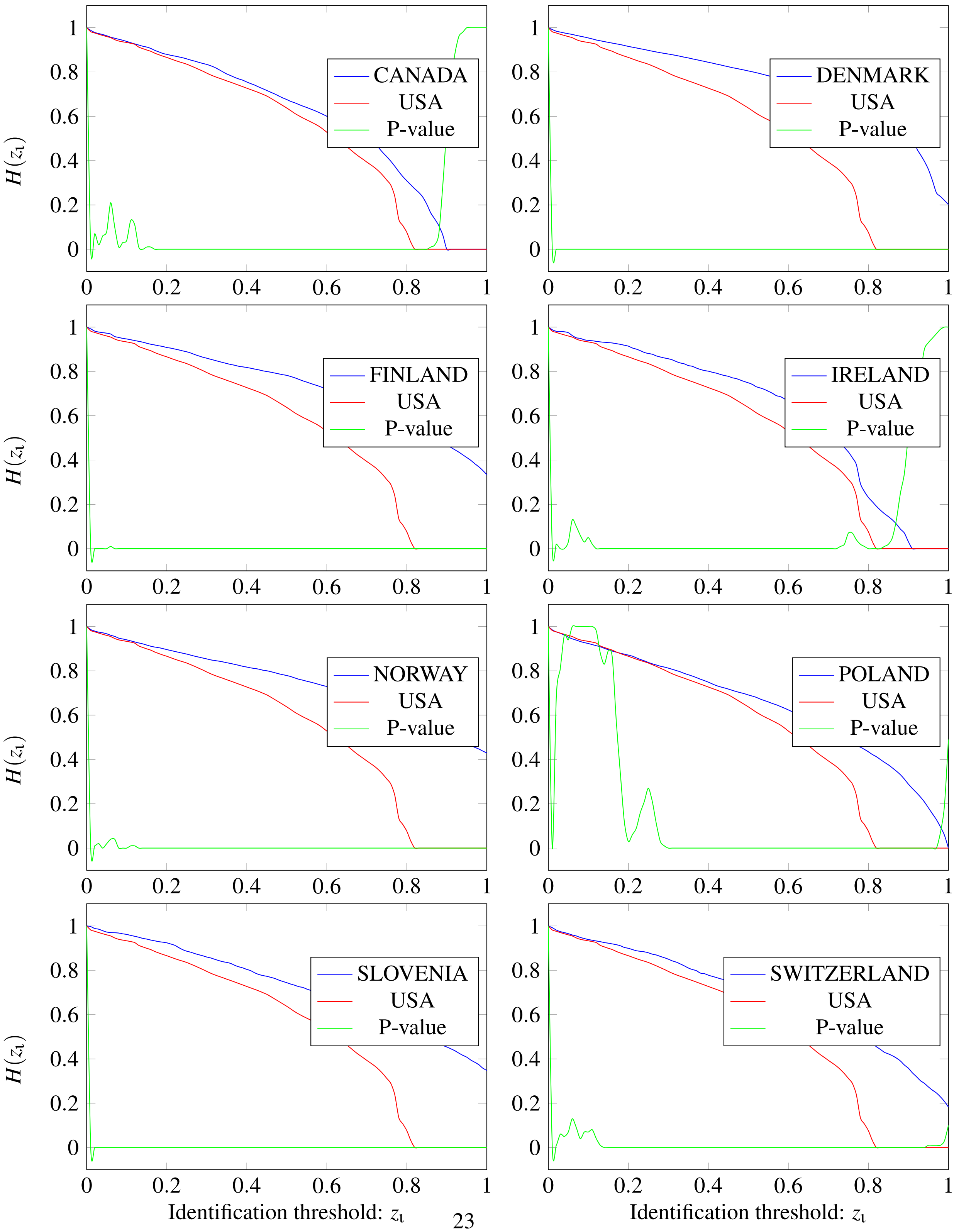
For any identification threshold lower than 0.6, the p-values of the test between Mexico and the USA equal 1, Figure 6. If we neglect the small values of identification thresholds close to zero, this test confirms that the USA does not stochastically dominate Mexico when the identification threshold is lower than 0.6. However, the USA stochastically dominates Mexico for an identification threshold interval in]0.6, 1[and p-values are close to zero in this interval. The same pattern is observed for the test between the USA and Peru where the USA does not dominate Peru for any identification threshold lower than 0.77. For these values of identification thresholds, the associated p-values of the tests equal 1. Nevertheless, the statistical test shows that the USA dominates Peru for some identification thresholds in]0.77, 1[where p-values equal zero.

Figure 6: Identification dominance curves and p-values for Mexico, Peru and the USA



In Figure 7, p-values of the tests between the USA and Denmark, Finland, Norway, Slovenia and Switzerland are always close to zero regardless of the identification threshold. This confirms that dominance is achieved everywhere between the USA and these countries. Whatever the identification threshold, these countries are more homogenous than the USA. Statistical tests confirm restricted dominance between the USA and Canada, Ireland and Poland where p-values curves are greater than 5% for some identification thresholds. The p-value curve between the USA and Canada goes above 0.05 when the threshold exceeds 0.85. Between the USA and Ireland, the p-value curve is above 0.05 when the threshold exceeds 0.82. Finally the p-value curve between the USA and Poland is above 0.05 for thresholds in]0, 0.27[.

Figure 7: Identification dominance curves and p-values for Canada, Denmark, Finland, Ireland, Norway, Poland, Slovenia, Switzerland and the USA



7 Conclusion

In this paper we develop a methodology to compare the homogeneity of two income distributions based on a class of polarization indices. The point of departure is an index of polarization where the alienation component is not taken into account. This constituent of a polarization index without alienation measures the within-group homogeneity of a distribution. It also represents the degree to which people feel alike and identified to each other in a distribution.

The theoretical part of the paper establishes first-order, second-order and higher-order stochastic dominance of homogeneity across two distributions. First-order stochastic dominance yields identification dominance curves drawn on an interval of identification thresholds. An income distribution is always more homogenous and has more identification, or its individuals feel more alike than an another income distribution, if its identification dominance curve is always above the one of the second distribution. An estimator of the identification dominance curve is derived and shown to be convergent. We use this property of convergence to derive the sampling distribution of the estimator. The methodology of Intersection Union tests is used to test whether dominance observed with dominance curves can be confirmed between two dominance curves.

The empirical part compares homogeneity pairwise in the income distributions of 11 countries. We find that scandinavian countries, namely Denmark, Finland and Norway, are more homogenous than the USA, Canada, Ireland, Mexico, Peru, Poland and Switzerland. Moreover, Mexico and Peru are less homogenous than the USA, which in turn is less homogenous than the rest of the countries. The p-values curves are drawn in the same graphs as dominance curves and allow to confirm dominance if any and to determine the interval of dominance. For Canada, Ireland, Mexico, Peru and Poland, dominance by the USA is achieved in restricted intervals of identification thresholds, while for the rest of the countries dominance is achieved everywhere.

Identification is low in the USA compared to other countries, but in other work we have found that alienation and polarization are high in the USA and low in scandinavian countries. As a consequence of these findings, the larger polarization in the USA is mostly attributed to large differences between incomes, and less to group sizes and the sense to which people feel identified to each other. Also, low polarization in scandinavian countries can be attributed to a high concentration of income within their population.

References

BERGER, R. L. (1982): "Multiparameter Hypothesis Testing and Acceptance Sampling," *Technometrics*, 24, 295–300.

——— (1996): "Likelihood Ratio Tests and Intersection-Union Tests," *Institute of Statistics Mimeo Series*.

BERGER, R. L. AND D. F. SINCLAIR (1984): "Testing Hypotheses concerning Unions of Linear Subspaces," *Journal of the American Statistical Association*, 79, 158–163.

DAVIDSON, R. AND J.-Y. DUCLOS (2000): "Statistical Inference for stochastic dominance and for the measurement of poverty and inequality," *Econometrica*, 68, 1435–1464.

———(2013): "Testing for Restricted Stochastic Dominance," *Econometric Reviews*, 32, 84–125.

DUCLOS, J. AND A. ARAAR (2006): *Poverty and Equity: Measurement, Policy and Estimation with DAD*, Economic Inequality and Poverty Series, Springer.

DUCLOS, J.-Y., J.-M. ESTEBAN, AND D. RAY (2004): "Polarization: Concepts, Measurement, Estimation," *Econometrica*, 72, 1737–1772.

ESTEBAN, J.-M. AND D. RAY (1994): "On the Measurement of Polarization," *Econometrica*, 62, 819–851.

———(2008): "Polarization, Fractionalization and Conflict," *Journal of Peace Research*, 45, 163–182.

FOSTER, J. AND A. SHORROCKS (1988): "Poverty orderings and welfare dominance," *Social Choice and Welfare*, 5, 179–198.

HOWES, S. (1993): "Asymptotic Properties of Four Fundamental Curves of Distributional Analisys," *Unpublished Paper, STICERD, London School of Economics*.

KAUR, A., B. PRAKASA RAO, AND H. SINGH (1994): "Testing for Second-Order Stochastic Dominance of Two Distributions," *Econometric Theory*, 10, 849–866.

LUXEMBOURG INCOME STUDY (LIS) DATABASE, http://www.lisdatacenter.org (multiple countries; October 2014). Luxembourg: LIS.

MCFADDEN, D. (1989): "Testing for Stochastic Dominance," in *Studies in the Economics of Uncertainty*, ed. by T. Fomby and T. Seo, Springer New York, 113–134.

Appendix

Consider a random sample of N independently and identically distributed observations of income $y_1, \cdots, y_N$ drawn from a distribution with cumulative distribution function F and an identification threshold z. For simplicity, let

$$\begin{aligned} H(z) &= \int_z^{\infty} dF_f(\phi) && (45) \\ &= \int I[\phi \geq z] dF_f(\phi) && (46) \end{aligned}$$

where ϕ is a density and F_f is the cumulative distribution function of the density f. An estimator of $H(z)$ is :

$$\begin{aligned} \hat{H}(z) &= \int I[\hat{\phi} \geq z] d\hat{F}_f(\phi) && (47) \\ &= N^{-1} \sum_{i=1}^{N} I[\hat{f}(y_i) \geq z] && (48) \\ &= \int I[\phi \geq z] d\hat{F}_f(\phi) \\ &+ \underbrace{N^{-1} \sum_{i=1}^{N} I[z - \hat{f}(y_i) + f(y_i) < f(y_i) < z]}_{E1} \\ &- \underbrace{N^{-1} \sum_{i=1}^{N} I[z < f(y_i) < z - \hat{f}(y_i) + f(y_i)]}_{E2} && (49) \end{aligned}$$

where $E1 - E2$ is an error term due to the fact that the density is estimated.

$$\begin{aligned} \hat{H}1 &= \int I[\phi \geq z] d\hat{F}_f(\phi) && (50) \\ &= \int I[\phi \geq z] d\underbrace{(\hat{F}_f - F_f)}_{O(N^{-\frac{1}{2}})}(\phi) + \int I[\phi \geq z] dF_f(\phi) && (51) \\ &\cong N^{-1} \sum I(f(y_i) \geq z) - 1 + F_f(z) + H(z) && (52) \end{aligned}$$

$$\begin{aligned}
E1 &= N^{-1}\sum_{i=1}^{N} I[z-\hat{f}(y_i)+f(y_i)<f(y_i)<z] \tag{53}\\
&= \int I[z-(\hat{\mathrm{v}}(v)-v)<v<z]d\hat{F}_f(v) \tag{54}\\
&= \int \underbrace{I[z-(\hat{\mathrm{v}}(v)-v)<v<z]}_{O(N^{-\frac{1}{2}})} d\underbrace{(\hat{F}_f-F_f)}_{O(N^{-\frac{1}{2}})}(v) + \int I\underbrace{[z-(\hat{\mathrm{v}}(v)-v)<v<z]}_{O(N^{-\frac{1}{2}})}dF_f(v) \tag{55}\\
&\cong O(N^{-1}) + \int I[z-(\hat{\mathrm{v}}(v)-v)<v<z]dF_f(v) \tag{56}\\
&\approx O(N^{-1}) + (\hat{\mathrm{v}}(z)-z)f_f(z)I[\hat{\mathrm{v}}(z)\geq z] \text{ using a Taylor expansion} \tag{57}\\
&\approx O(N^{-1}) + (\hat{\mathrm{v}}(z)-z)_+f_f(z) \text{ where } x_+=max(0,x) \tag{58}
\end{aligned}$$

$$\begin{aligned}
E2 &= N^{-1}\sum_{i=1}^{N} I[z<f(y_i)<z-\hat{f}(y_i)+f(y_i)] \tag{59}\\
&= \int I[z<v<z-\hat{\mathrm{v}}(v)+v]d\hat{F}_f(v) \tag{60}\\
&= \int \underbrace{I[z<v<z-\hat{\mathrm{v}}(v)+v]}_{O(N^{-\frac{1}{2}})} d\underbrace{(\hat{F}_f-F_f)}_{O(N^{-\frac{1}{2}})}(v) + \int I\underbrace{[z<v<z-\hat{\mathrm{v}}(v)+v]}_{O(N^{-\frac{1}{2}})}dF_f(v) \tag{61}\\
&\cong O(N^{-1}) + \int I[z<v<z-\hat{\mathrm{v}}(v)+v]dF_f(v) \tag{62}\\
&\approx O(N^{-1}) - (\hat{\mathrm{v}}(z)-z)f_f(z)I[\hat{\mathrm{v}}(z)<z] \text{ using a Taylor expansion} \tag{63}\\
&\approx O(N^{-1}) - (\hat{\mathrm{v}}(z)-z)_-f_f(z) \text{ where } x_-=min(0,x) \tag{64}
\end{aligned}$$

$$\begin{aligned}
\hat{H}(z) &\cong N^{-1}\sum I(f(y_i)\geq z)-1+F_f(z)+H(z)+0(N^{-1})+(\hat{\mathrm{v}}(z)-z)_+f_f(z)\\
&- 0(N^{-1})+(\hat{\mathrm{v}}-z)_-f_f(z)\\
&\approx N^{-1}\sum I(f(y_i)\geq z)-1+F_f(z)+H(z)+[(\hat{\mathrm{v}}(z)-z)_+ + (\hat{\mathrm{v}}(z)-z)_-]f_f(z)\\
&\equiv N^{-1}\sum I(f(y_i)\geq z)-1+F_f(z)+H(z)+(\hat{\mathrm{v}}(z)-z)f_f(z) \tag{65}
\end{aligned}$$

Then

$$\begin{aligned}
N^{\frac{1}{2}}(\hat{H}(z)-H(z)) &= N^{-\frac{1}{2}}\sum I(f(y_i)\geq z)+N^{\frac{1}{2}}(-1+F_f(z))+N^{\frac{1}{2}}(\hat{\mathrm{v}}(z)-z)f_f(z) \tag{66}\\
&= N^{-\frac{1}{2}}\sum I(f(y_i)\geq z)+N^{\frac{1}{2}}(-1+F_f(z))+N^{\frac{1}{2}}\left(\frac{N^{-1}}{h}\sum k(y_i,f^{-1}(z))-z\right)f_f(z)
\end{aligned}$$

where $k(x,y)=k(\frac{x-y}{h})$ is a kernel function and h a bandwidth. However, $f^{-1}(z)$ does not always correspond to a unique value of income.

Let $(y_1^*, y_2^*, \cdots, y_J^*)$ be J values of income for which $f(y_j^*) = z$.

$$(\hat{\mathrm{v}}(z) - z) f_f(z) = J^{-1}\left[\sum_{j=1}^{J} (\hat{\mathrm{v}}_j(z) - z) f_{f,j}(z)\right] \tag{67}$$

where

$$\begin{aligned}
\hat{\mathrm{v}}_j(z) &= \frac{N^{-1}}{h}\sum_{i=1}^{N} k_p(y_i, f_{f,j}^{-1}(z)) && (68)\\
&= \frac{N^{-1}}{h}\sum_{i=1}^{N} k_p(y_i, y_j^*) && (69)\\
&\equiv \hat{f}(y_j^*) && (70)
\end{aligned}$$

and $f_{f,j}(z)$ verifies the following relation:

$$f_f(z) = J^{-1}\sum_{j=1}^{J} f_{f,j}(z). \tag{71}$$

Indeed, an estimator of the density of densities is:

$$\begin{aligned}
\hat{f}_f(z) &= \frac{N^{-1}}{h}\sum_{i=1}^{N} k_f(f(y_i), f = z) && (72)\\
&= N^{-1}\sum_{i=1}^{N} J^{-1}\sum_{j=1}^{J} \frac{k_p(y_i, y_j^*)}{J^{-1}\sum_{j=1}^{J} k_p(y_i, y_j^*)} k_f(f(y_i), f = z) && (73)\\
&= J^{-1}\sum_{j=1}^{J} \underbrace{N^{-1}\sum_{i=1}^{N} \frac{k_p(y_i, y_j^*)}{J^{-1}\sum_{j=1}^{J} k_p(y_i, y_j^*)} k_f(f(y_i), f = z)}_{f_{f,j}(z)} && (74)
\end{aligned}$$

where $k_p(.,.)$ is a kernel function defined in the income space and $k_f(.,.)$ the one defined in the density space.

$$\hat{\mathrm{v}}_j(z) - z = \frac{N^{-1}}{h}\sum_{i=1}^{N} (k_p(y_i, y_j^*) - hz) \tag{75}$$

with $E(\hat{\mathrm{v}}_j(z) - z) = 0$. Then,

$$\begin{aligned}
N^{\frac{1}{2}}(\hat{H}(z) - H(z)) &= N^{-\frac{1}{2}} \sum I(f(y_i) \geq z) + N^{\frac{1}{2}}(-1 + F_f(z)) && (76)\\
&+ N^{\frac{1}{2}} J^{-1}\sum_{j=1}^{J}\left(\frac{N^{-1}}{h}\sum_{i=1}^{N}(k_p(y_i, y_j^*) - hz)\right) f_{f,j}(z)\\
&= N^{-\frac{1}{2}} \sum I(f(y_i) \geq z) + N^{\frac{1}{2}}(-1 + F_f(z)) && (77)\\
&+ \frac{N^{-\frac{1}{2}} J^{-1}}{h}\sum_{j=1}^{J}\left(\sum_{i=1}^{N}(k_p(y_i, y_j^*) f_{f,j}(z)\right) - N^{\frac{1}{2}} z f_f(z)
\end{aligned}$$

It follows that $E[N^{\frac{1}{2}}(\hat{H}(z)-H(z))]=0$ and we derive that $N^{\frac{1}{2}}(\hat{H}(z)-H(z))$ *is asymptotically normal with mean zero.*

Variance of $\hat{H}(z)$

From equation (77), we have

$$\begin{aligned} \hat{H}(z)-H(z) &= N^{-1}\sum I(f(y_i)\geq z)+(-1+F_f(z)) \qquad (78)\\ &+ J^{-1}\sum_{j=1}^{J}\left(\frac{N^{-1}}{h}\sum_{i=1}^{N}(k_p(y_i,y_j^*)f_{f,j}(z)\right)-zf_f(z) \end{aligned}$$

such that

$$\begin{aligned} Var(\hat{H}(z)-H(z)) &= Var\left[N^{-1}\sum I(f(y_i)\geq z)+J^{-1}\sum_{j=1}^{J}\left(\frac{N^{-1}}{h}\sum_{i=1}^{N}(k_p(y_i,y_j^*)f_{f,j}(z)\right)\right] \qquad (79)\\ &= Var\left[N^{-1}\sum I(f(y_i)\geq z)+J^{-1}\sum_{j=1}^{J}\hat{v}_j(z)f_{f,j}(z)\right] \qquad (80)\\ &= Var\left[N^{-1}\sum I(f(y_i)\geq z)+\Phi(\theta_1(z),\cdots,\theta_J(z))\right] \qquad (81)\\ &= N^{-1}(1-F_f(z))F_f(z)+Var(\Phi)+2\mathrm{cov}\left(N^{-1}\sum I(f(y_i)\geq z),\Phi\right) \end{aligned}$$

where $\theta_j(z)=\hat{v}_j(z)f_{f,j}(z)=\hat{f}(y_j^*)f_{f,j}(z)$ and $\Phi=\Phi(\theta_1(z),\cdots,\theta_J(z))=J^{-1}\sum_{j=1}^{J}\hat{v}_j(z)f_{f,j}(z)$

$$\begin{aligned} Var(\theta_j(z)) &= Var\left(f_{f,j}(z)\frac{N^{-1}}{h}\sum_{i=1}^{N}k_p(y_i,y_j^*)\right) \qquad (82)\\ &= f_{f,j}^2(z)Var(\hat{f}(y_j^*)) \qquad (83)\\ &= f_{f,j}^2(z)\frac{N^{-1}}{h}f(y_j^*)\int k_p^2(\psi_j)d\psi_j, \qquad (84) \end{aligned}$$

where $\psi_j=\frac{y-y_j^*}{h}$. Let V be the vector of the first derivatives of Φ with respect to $\theta_j, j=1,\cdots J$ and M be the matrix of variance/covariance of (θ_j):

$$V = \begin{pmatrix} \frac{\partial\Phi}{\theta_1} \\ \frac{\partial\Phi}{\theta_2} \\ \vdots \\ \frac{\partial\Phi}{\theta_J} \end{pmatrix} = J^{-1}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad (85)$$

$$
\begin{aligned}
Cov(\theta_j(z),\theta_k(z)) &= Cov\left[\frac{N^{-1}}{h}\sum_{i=1}^{N}(k_p(y_i,y_j^*)f_{f,j}(z),\frac{N^{-1}}{h}\sum_{t=1}^{N}(k_p(y_t,y_k^*)f_{f,k}(z)\right] && (86)\\
&= f_{f,j}(z)f_{f,k}(z)Cov\left[\hat{f}(y_j^*),\hat{f}(y_k^*)\right] && (87)\\
&= \begin{cases} Var(\theta_j(z))\ \ if\ \ j=k \\ 0\ \ if\ \ j\neq k. \end{cases} && (88)
\end{aligned}
$$

.

$$
M = \begin{pmatrix} \mathrm{var}(\theta_1) & 0 & \cdots & 0 \\ 0 & \mathrm{var}(\theta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{var}(\theta_J) \end{pmatrix} \qquad (89)
$$

Using Rao's (1973) approach, the variance of Φ is given by:

$$
\mathrm{Var}(\Phi) = V'MV = J^{-2}\sum_{j=1}^{J}\mathrm{var}(\theta_j) \qquad (90)
$$

where V' is the transpose of V.

$$
\begin{aligned}
\mathrm{cov}(N^{-1}\sum_{i=1}^{N}I(f(y_i)\geq z),\Phi) &= \mathrm{cov}\left(N^{-1}\sum_{i=1}^{N}I(f(y_i)\geq z),J^{-1}\sum_{j=1}^{J}\theta_j\right) && (91)\\
&= N^{-1}J^{-1}\sum_{i}^{N}\sum_{j}^{J}\mathrm{cov}(I(f(y_i)\geq z),\theta_j) && (92)\\
&= J^{-1}\sum_{j}^{J}[E(\theta_j|f(y)\geq z)(1-F_f(z))-(1-F_f(z))E(\theta_j)]
\end{aligned}
$$

$$
E(\theta_j|f(y)\geq z) = f_{f,j}(z)E(\hat{f}(y_j^*)|f(y)\geq z) \qquad (93)
$$

$$
\begin{aligned}
E(\theta_j) &= f_{f,j}(z)E(\hat{f}(y_j^*)) && (94)\\
&= f_{f,j}(z)\int k_p(\psi_j)f(h\psi_j+y_j^*)d\psi_j && (95)
\end{aligned}
$$

Finally

$$
\begin{aligned}
\lim_{N\to\infty}NVar(\hat{H}(z)-H(z)) &= (1-F_f(z))F_f(z)+J^{-2}\sum_{j=1}^{J}\mathrm{var}(\theta_j) && (96)\\
&+ 2J^{-1}(1-F_f(z))\sum_{j}^{J}[E(\theta_j|f(y)\geq z)-E(\theta_j)].
\end{aligned}
$$