Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2011-18

Measuring Segregation on Small Units : A Partial Identification Analysis

X. D'HAULTFOEUILLE[1]
R. RATHELOT[2]

[1] CREST, Xavier.dhaultfoeuille@ensae.fr
[2] CREST, roland.rathelot@ensae.fr

Measuring Segregation on Small Units: A Partial Identification Analysis*

Xavier D'Haultfœuille[†] and Roland Rathelot[‡]

May 2011

Abstract

We consider the issue of measuring segregation in a population of small units, such as small firms or classrooms. Segregation is defined as an inequality index on the (random) probability that an individual of a given unit belongs to the minority group. Because this probability is measured with error by a proportion, standard estimators are inconsistent. Moreover, the corrections considered previously in the literature are valid only under restrictive conditions. We model this problem as a binomial mixture and show that under this testable assumption, only the first moments of the underlying probability are identified. As a result, segregation indices are only partially identified in general. Under conditions satisfied by standard segregation indices, we show that the sharp bounds of the identification region can be easily obtained by an optimization over a low dimensional space. We also develop inference on these bounds, by providing confidence intervals on the true parameter and the identification interval and considering tests of the binomial mixture model. Finally, we apply our framework to measure the segregation of immigrants in small French firms.

JEL Classification: C13, C14 and J71.

Keywords: segregation, small units, partial identification.

[*] We would like to thank Romain Aeberhardt and Mathias André for their comments. We also benefited from discussions with Christian Bontemps, Guy Laroque, Adam Rosen and the participants of the 21st EC^2 conference in Toulouse, Crest-LMI and INSEE-D3E seminars.

[†] CREST (INSEE), xavier.dhaultfoeuille@ensae.fr.

[‡] CREST, roland.rathelot@ensae.fr.

1 Introduction

Being able to measure the degree of segregation of a population across units is a crucial step to understand a phenomenon and design adequate policies. In several cases, however, the very nature of the phenomenon under study makes it difficult to proceed to measurement. In particular, when the units contain few individuals, the usual segregation indices prove to be poor estimates of the actual level of segregation, an issue known as the *small-unit bias*. Social scientists and economists studying workplace segregation or school segregation experience it as an everyday issue, as an important proportion of firms have less than ten employees and classrooms are around 20 pupils.[1]

When the number of individuals per unit is small, the observed proportion of a minority group in the unit becomes a poor estimate of the true unobserved probability that a given individual belongs to the minority group. Intuitively, if the unit is divided by two, the index computed using these proportions is going to be higher, not because a higher segregation takes place, but because of the noise induced on the measurement of the proportions. As Cortese et al. (1976) made it clear, naive segregation indices measure, in this case, the distance to *evenness*, when observed proportions are all equal across groups, while one would rather be interested in the distance to *randomness*, which corresponds to the situation where the true unobserved probability is constant.

Several works propose solutions to deal with this issue. The most common way is to provide corrected versions of the indices, in an attempt to extract the signal from the noise. Winship (1977) has been the first to propose a corrected Duncan index. Carrington & Troske (1997) developed his idea and have also proposed an adjusted index, close to Winship's. Allen et al. (2009) proposed a correction based on bootstrap. Finally, Rathelot (2011) develops an estimator which is consistent under a parametric condition on the underlying distribution. He also shows that his index and the one proposed by Allen *et al.* perform better than other solutions for many standard distributions on the underlying probability. However, none of these approaches is consistent for any distribution on this probability.

In this paper, we first reconsider the problem from an identification viewpoint. In line with the literature, we suppose that the observed number of people belonging to the minority

[1] Measuring workplace segregation at the level of the firm has been recently featured in Carrington & Troske (1995), Carrington & Troske (1998*a*), Carrington & Troske (1998*b*), Bayard et al. (1999) or Hellerstein & Neumark (2008). Likewise, there are recent attempts to measure segregation at the level of the schools or the classrooms; see Allen et al. (2009) or Söderström & Uusitalo (2010).

group in a given unit follows, conditional on the unobserved probability, a binomial distribution. On the other hand, we remain completely agnostic on the distribution of the unobserved probability. The binomial assumption, which we show is testable, allows us to identify the first K moments of the distribution of the unobserved probability, where K denotes the size of the units. Because most of the existing segregation indices depend on the whole distribution of this probability, not only on its first moments, these indices are only partially identified in general. Bounds can be obtained by minimizing or maximizing these indices over distributions whose first moments match those identified in the data. This problem is though a difficult one, as the space of corresponding distributions is of infinite dimension in general.

Another contribution of this paper is to prove that under a mild convexity condition satisfied by most segregation indices, the upper bound can be obtained by maximizing over discrete distributions with at most $K+1$ points of support only. The lower bound follows similarly if the unknown parameter is linear in the distribution, a condition satisfied for instance by the popular Duncan and Theil indices. Interestingly, this result is related to the one of Chernozhukov et al. (2009) in the rather different framework of nonlinear panel data. In such models, bounds on marginal effects can be obtained by maximizing some functionals over the distribution of the fixed effect. Similarly to us, they show that it suffices actually to restrict oneself to discrete distributions with a low number of support points.

We also develop inference on these bounds, using a two-step procedure. In the first step, we estimate the vector of moments by GMM under the constraint that it belongs to the set of moments of distributions on $[0,1]$. The estimator takes a closed and very simple form whenever the constraint is slack. Otherwise, we use a characterization of the projection on this set to transform this problem into an optimization under only linear equality and inequality constraints. In the second step, the bounds are estimated by optimizing over finite-dimensional distributions whose first moments match the first-step estimator. Interestingly, the lower and upper bounds coincide when the constraint on the vector of moment is binding, and no optimization is needed in this case. The estimated bounds are asymptotically normal, with easy to compute asymptotic variances, leading to a simple confidence interval which is asymptotically valid for the true parameter. We also propose a conservative confidence interval on the true identification interval. Finally, we develop tests on the binomial mixture model, based on the distance between the unconstrained estimated vector of moment and the constrained one.

Monte Carlo simulations indicate that our method works well for finite samples. They also show that even for modest unit sizes ($K = 8$, typically), the constraint on the vector of moment is binding most of the times for sample sizes as large as $10,000$, leading in most cases to an estimated identification interval reduced to a single point. More generally, the length of the confidence intervals mostly stems from sampling variation, not partial identification for typical unit and sample sizes.

Finally, we apply our framework to measure the segregation of immigrants in small French firms, conditional on plant size. The first interesting result is that our method proves to work well is this context: when the plant size is larger than 2 or 3, the identification region is already informative. Over the whole curve, Non-European workers seem to be more segregated than European (foreign) workers, even though the difference between the Theil indices across groups never exceeds 10 points. Contrary to what is suggested by the naive or Allen et al. (2009) estimator, we cannot reject at standard levels that there is no relationship between plant size and segregation. Besidesy, the methods introduced by Carrington & Troske (1997), Allen et al. (2009) and Rathelot (2011) provide estimates that are always outside the identification region for plant sizes larger than 4, and even often outside the corresponding confidence intervals.

The paper is organized as follows. Section two presents the binomial mixture model and studies partial identification of parameters of interest in this model. Section three develops inference on the bounds and a test of the binomial assumption. Monte Carlo simulations are drawn in Section four, and the application to segregation in the workplace is displayed in the last section. All proofs are deferred to the appendix.

2 Identification

2.1 The binomial mixture model

The population is assumed to be split into two groups, a group of interest and the rest of the population, and to be distributed across units. Units may represent geographical areas, classrooms, or, as in our application, firms. For simplicity, we consider the size of the units to be constant and equal to K. Equivalently, our analysis can be seen as conditional on K. Now assume that there exists a random variable p taking values in $[0, 1]$ that represents the probability for any individual belonging to a given unit to be a member of the population of interest. We focus on the measure of segregation of the population of interest across units, which may be defined by a real parameter $\theta = g(F_p)$, g being a functional defined on the set $\mathcal{D}$ of distribution functions on $[0, 1]$ and F_p being the distribution function of

p. We also suppose hereafter that g is continuous with respect to the metric induced by the supremum norm. Standard segregation indices include the Gini, the dissimilarity (or Duncan) and the Theil indices but our identification results below will be more general.

The main problem in the inference on θ is that p is not directly observed. We only observe the size of the unit, K, and the number of individuals in that unit that belong to the group of interest, X (we consider hereafter the case where we only observe K and a subsample of size $n_K < K$ of individuals in the unit). We posit that individuals are selected into units independently from each others in terms of their membership of the group of interest. In this case, X follows, conditional on p, a binomial distribution $B(K, p)$. Because p is random and unobserved, this model is called a binomial mixture (see, e.g., Lord, 1969, Wood, 1999). Note that the independence condition may not hold. The presence of an immigrant in a firm may, for instance, increase the probability that another immigrant is employed in this firm. However, in the absence of detailed data on the selection process into units, this seems to us to be the most transparent assumption. It is also assumed by Carrington & Troske (1997), Allen et al. (2009) or Rathelot (2011), among others. Finally, as we shall see below, this assumption is testable.

Intuitively, since the distribution of X is defined by K probabilities, namely $P = (P_1, ..., P_K)'$ (where $P_i = \Pr(X = i)$), we expect it to convey information on $K + 1$ parameters on F_p. Actually, because $P_K = E(p^K)$ and

$$\begin{aligned} P_{K-k} &= \binom{K}{k} E\left[p^{K-k}(1-p)^k\right] \\ &= \binom{K}{k}\left\{E(p^{K-k}) + \sum_{j=1}^{k}\binom{k}{j}(-1)^j E(p^{K-k+j})\right\}, \end{aligned} \tag{2.1}$$

an immediate backward induction shows that all moments of p of order up to K are identified. However, these K moments do not correspond in general to a unique distribution of p. This is the reason why θ is only partially identified in general.

We may not observe all individuals in the unit. A common situation is indeed that $n_K < K$ individuals only are sampled from the unit. In this case, X denotes the number of individuals belonging to the group of interest in this subsample. As previously, X follows, conditional on p (and n_K if it is random), a binomial distribution $B(n_K, p)$. Hence, the result above applies by replacing K by n_K. The main difference in this case is that, if n_K is random, it is plausible to assume it to be independent of p conditional on K, whereas the assumption that p is independent of K is stronger in general. Under this condition, it is easy to see that the $\overline{n}_K$ first moments of p are identified, where $\overline{n}_K$ denotes the maximum of the support of n_K. For the sake of clarity, we continue to suppose that all individuals in the units are sampled, so that $n_K = K$, although the results below apply directly to the case where $n_K < K$.

The binomial model allows us to recover the vector $m_0 = (m_{01}, ..., m_{0K})'$ of the K first raw moments of p. Such a vector should satisfy some restrictions, such as variance positivity ($m_{02} \geq m_{01}^2$). These restrictions may not hold if the model is not binomial, making it testable. For instance, supposing that $K = 2$, the vector $P = (0.6, 0.3)'$ should correspond to the vector of raw moments $m_0 = (0.6, 0.3)'$ according to the binomial model. But $0.3 - 0.6^2 < 0$, which violates the restriction that a variance is positive. In other words, this vector P invalidates the binomial mixture model.

More generally, the issue of whether a given vector belongs to the set $\mathcal{M}$ of first K moments of a probability distribution on $[0, 1]$ is known as the truncated Hausdorff problem. Several necessary and sufficient conditions have been established for this problem (see, e.g., Krein & Nudel'man, 1977, Curto & Fialkow, 1991). Proposition 2.1, which is proved for instance by Krein & Nudel'man (1977, Theorem III.2.4), provides one that is rather simple to compute. We let afterwards L denote the integer part of $K/2$, and, for a given vector $\mu = (\mu_1, ..., \mu_K)'$, A_μ, B_μ, C_μ denote the square matrices of size $L + 1$, $L + 1$ and L respectively, with typical (i, j) term equal to μ_{i+j-2}, μ_{i+j-1} and $\mu_{i+j-1} - \mu_{i+j}$ respectively (where we let $\mu_0 = 1$).

Proposition 2.1 *When K is odd (resp. even), $\mu \in \mathcal{M}$ if and only if $A_\mu - B_\mu$ and B_μ (resp. A_μ and C_μ) are positive.*

2.2 Bounds on segregation indices

Let $\mathcal{P}_{m_0} = \{F \in \mathcal{D} : \int x^k dF(x) = m_{0k} \text{ for all } k = 1...K\}$ denote the set of distributions on $[0, 1]$ that match the identified moments of p. The lower and upper bounds $\underline{\theta}_0$ and $\overline{\theta}_0$ of the identified set of θ_0 satisfy

$$\underline{\theta}_0 = \inf_{F \in \mathcal{P}_{m_0}} g(F), \quad \overline{\theta}_0 = \sup_{F \in \mathcal{P}_{m_0}} g(F). \tag{2.2}$$

For any $\varepsilon > 0$, let $\underline{F}_\varepsilon$ (resp. $\overline{F}_\varepsilon$) be such that $g(\underline{F}_\varepsilon) < \underline{\theta}_0 + \varepsilon$ (resp. $g(\overline{F}_\varepsilon) > \overline{\theta}_0 - \varepsilon$). Then, by the intermediate value theorem applied to the continuous function $t \mapsto g(t\underline{F}_\varepsilon + (1-t)\overline{F}_\varepsilon)$, the interval $[\underline{\theta}_0 + \varepsilon; \overline{\theta}_0 - \varepsilon]$ is included in the identified set of θ_0. Because ε was arbitrary, this identified set is thus the whole interval $[\underline{\theta}_0, \overline{\theta}_0]$.

Computing the bounds with (2.2) is not convenient as it involves finding an infimum and a supremum of a function over an infinite dimensional set. The idea we develop here is to restrict oneself to distributions with finite supports by considering

$$\underline{\theta}_0^k = \inf_{F \in \mathcal{P}_{m_0}^k} g(F), \quad \overline{\theta}_0^k = \sup_{F \in \mathcal{P}_{m_0}^k} g(F),$$

where $\mathcal{P}_{m_0}^k$ denotes the subset of $\mathcal{P}_{m_0}$ with at most k points of support. Of course, because the optimization set is smaller, we only obtain inner bounds in general, i.e. $\underline{\theta}_0^k \geq \underline{\theta}_0$ and $\overline{\theta}_0^k \leq \overline{\theta}_0$. However, Theorem 2.2 below establishes three useful results on these inner bounds.

Theorem 2.2 *1.* $\lim_{k\to\infty} \underline{\theta}_0^k = \underline{\theta}_0$ *and* $\lim_{k\to\infty} \overline{\theta}_0^k = \overline{\theta}_0$*;*

2. If g is convex, $\overline{\theta}_0^{K+1} = \overline{\theta}_0$*;*

3. If g is linear, $\underline{\theta}_0^{K+1} = \underline{\theta}_0$ *and* $\overline{\theta}_0^{K+1} = \overline{\theta}_0$*.*

Part 1 shows that the inner bounds tend to the true bounds when $k \to \infty$. Intuitively, this stems from the continuity of g and the fact that discrete distributions are dense in $\mathcal{D}$. However, this is not as straightforward as it might seem, because discrete distributions in $\mathcal{P}_{m_0}$ may have not been dense in $\mathcal{P}_{m_0}$.[2] Part 3 is actually a corollary of Part 2, when applied to g and $-g$. The intuition of Part 2 can be explained as follows. First, using Part 1, we can reach the supremum $\overline{\theta}_0$, up to an arbitrary small constant, by computing the maximum of $g(F)$ over $\mathcal{P}_{m_0}^{k_0}$, for k_0 large enough. Second, by the convexity of g, Minkowski's and Carathéodory's theorems (see, e.g., Hiriart-Urruty & Lemaréchal, 2001, Theorems 2.3.4 and 1.3.6 respectively), we show that the maximum over this set is reached by at least one extremal element of $\mathcal{P}_{m_0}^{k_0}$. Using Carathéodory's theorem once more, we finally prove that the extremal elements of $\mathcal{P}_{m_0}^{k_0}$ actually have at most $K+1$ points of support. Note that $\mathcal{P}_{m_0}^{K+1}$ can be seen as a subset of $[0,1]^{2(K+1)}$, as any $F \in \mathcal{P}_{m_0}^{K+1}$ is defined by its support points and associated probabilities. This makes the maximization rather straightforward in practice.

Theorem 2.2 is fortunate because the functionals involved in standard segregation indices are either linear or convex. Because $E(p)$ is identified, The Duncan index $E|p - E(p)|/(2E(p)(1-E(p)))$, for instance, is a known function of $g_D(F) = \int |x - E(p)| dF(x)$, which is linear in F. Similarly, the Theil index $1 - E(p\ln(p)/[E(p)\ln(E(p))]$ involves only the linear functional $g_T(F) = \int x \ln(x) dF(x)$. Finally, the Gini coefficient $(1 - E(p) - \int F_p^2(x)dx)/E(p)$ is a known function of $g_G(F) = \int F^2(x)dx$, which is a convex functional.

[2] Technically, and letting $\overline{A}$ denote the closure of any set A, we only have, for any sets A and B, $\overline{A \cap B} \subset \overline{A} \cap \overline{B}$. Thus, even if $\overline{\cup_{k=1}^{\infty} \mathcal{D}^k} = \mathcal{D}$, where $\mathcal{D}^k$ denotes the set of distributions with at most k support points, we only have $\overline{\cup_{k=1}^{\infty} \mathcal{P}_{m_0}^k} = \overline{(\cup_{k=1}^{\infty} \mathcal{D}^k) \cap \mathcal{P}_{m_0}} \subset \overline{\cup_{k=1}^{\infty} \mathcal{D}^k} \cap \overline{\mathcal{P}_{m_0}} = \mathcal{P}_{m_0}$, and the other inclusion is not trivial.

2.3 Comparison with other approaches

Broadly speaking, other approaches have ignored so far the issue of partial identification that we have shown to arise here. Rather, they focus on the estimation of parameters that are identified, but different from θ_0 in general. The first and perhaps most natural possibility is to ignore the randomness due to the small size of the unit, and make as if $X = Kp$. This amounts to estimating the parameter $\widetilde{\theta} = g(F_{\frac{X}{K}})$. However, if g is monotonous with respect to the second-order dominance, as is the case of most inequality indices (including the three considered above), this parameter is always greater than $\overline{\theta}_0$.

Proposition 2.3 *Suppose that g is decreasing with respect to the second-order dominance. Then $\widetilde{\theta} \geq \overline{\theta}_0$. Moreover, the inequality is strict if g is strictly decreasing*[3] *and the support of p is not reduced to* $\{0, 1\}$.

Previous approaches have soon recognized this small-unit bias. The most commonly used method to correct for it is the one introduced by Carrington & Troske (1997), which is based on earlier works by Winship (1977) and Cortese et al. (1978). Suppose here, and without loss of generality if g is bounded, that g ranges from 0 to 1, the corrected index θ_{CT} of Carrington & Troske (1997) is defined by

$$\theta_{CT} = \frac{\widetilde{\theta} - \theta^*}{1 - \theta^*},$$

where $\theta^* = g(F_{\frac{X^*}{K}})$ (where $X^* \sim B(E(p), K)$) is the naive parameter that would be obtained if all units had the same probability, that is if segregation was zero. The index θ_{CT} can be seen as an affine correction that is valid in the two polar cases where there is no segregation (because $\theta_{CT} = \widetilde{\theta} = \theta_0 = 0$ in this case) or if segregation is maximal (because then $\theta_{CT} = \theta_0 = 1$). However, in general θ_{CT} does not lie inside the interval $[\underline{\theta}_0, \overline{\theta}_0]$, as Figure 1 below shows.

Allen et al. (2009) propose a bootstrap correction of the segregation index. Their idea is that we can obtain a good approximation of the discrepancy between $\widetilde{\theta} = g(F_{\frac{X}{K}})$ and θ_0 by bootstrap, and then correct for this discrepancy. Namely, they propose to approximate this discrepancy by $\widetilde{\widetilde{\theta}} - \widetilde{\theta}$, where $\widetilde{\widetilde{\theta}} = g(F_{\frac{X^{**}}{K}})$ and $X^{**}|X \sim B(K, \frac{X}{K})$. The corrected index is then:

$$\theta_{ABW} = 2\widetilde{\theta} - \widetilde{\widetilde{\theta}} \; (= \widetilde{\theta} + \widetilde{\theta} - \widetilde{\widetilde{\theta}}).$$

[3] Here we say that g is strictly decreasing with respect to the second-order dominance if, whenever $\int u(x)dF(x) > \int u(x)dG(x)$ for all strictly concave u, we have $g(F) > g(G)$.

The idea behind this parameter is that, if $\frac{X}{K}$ was distributed as p, then we would have $\widetilde{\theta}-\theta_0 = \widetilde{\widetilde{\theta}}-\widetilde{\theta}$. Formally, one can show that in some circumstances, θ_{ABW} reduces the order of the bias of $\widetilde{\theta}$. Suppose that $g(F) = \int \gamma(x)dF$, where γ is twice continuously differentiable and $\gamma(E(p)) = 0$. Such restrictions are satisfied for the Theil index, for instance. Then, by a second order Taylor expansion and using decompositions of variances, we get

$$\begin{aligned}
\theta_0 &\simeq \frac{\gamma''(E(p))}{2}V(p) \\
\widetilde{\theta} &\simeq \frac{\gamma''(E(p))}{2}V\left(\frac{X}{K}\right) = \frac{\gamma''(E(p))}{2}\left[V(p) + \frac{E(p(1-p))}{K}\right] \\
\widetilde{\widetilde{\theta}} &\simeq \frac{\gamma''(E(p))}{2}V\left(\frac{X^{**}}{K}\right) = \frac{\gamma''(E(p))}{2}\left[V(p) + \frac{E(p(1-p))}{K}\left(2-\frac{1}{K}\right)\right].
\end{aligned}$$

This suggests that the leading term in $\widetilde{\theta}-\theta_0$, when considering that $K \to \infty$, is $\gamma''(E(p))E(p(1-p))/K$. This term disappears in $\theta_{ABW} - \theta_0$, so that we expect the bias to be of smaller order (namely $1/K^{3/2}$ or $1/K^2$). In the simulations run by Rathelot (2011), it seems that θ_{ABW} performs better when the true level of segregation is high.

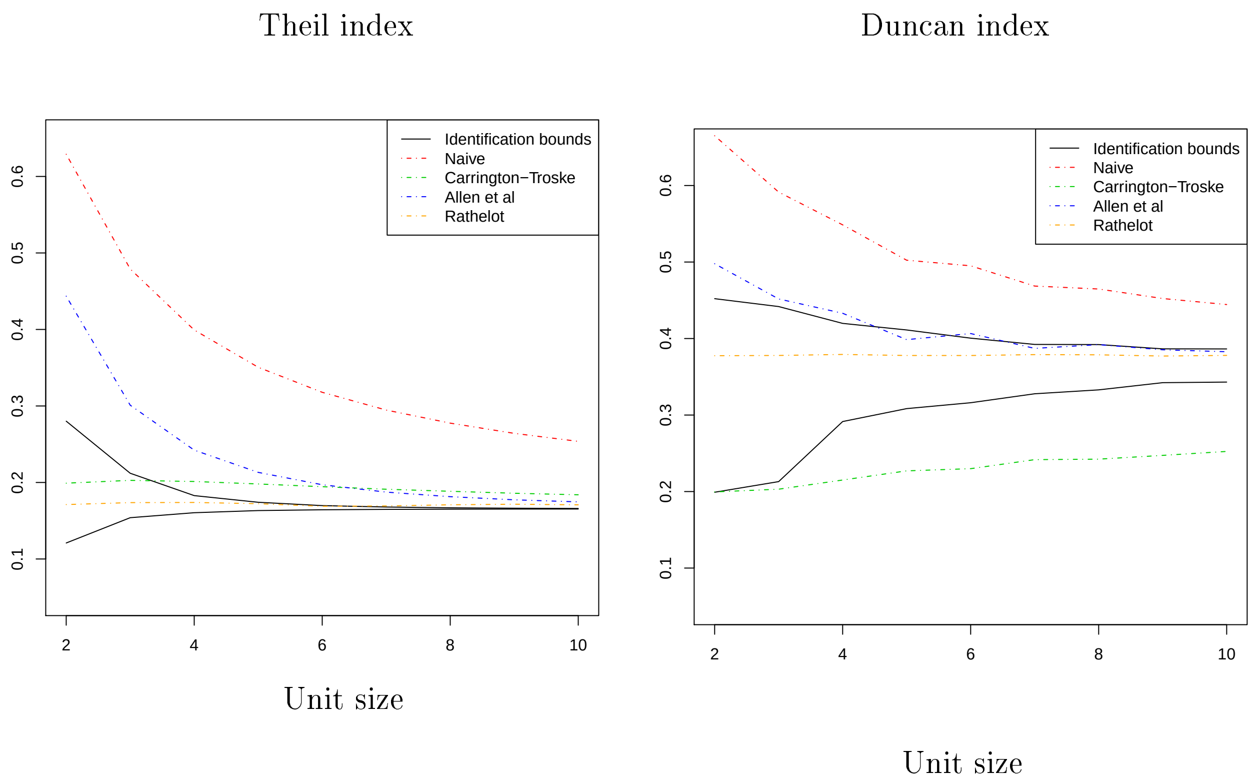
Finally, Rathelot (2011) introduces another correction based on the parametric assumption that the distribution of p is a mixture of two beta distributions. Once the five parameters have been obtained, the segregation indices may easily be deduced. On the other hand, these indices do not lie inside the identification interval whenever the first moments of p do not correspond to those of a mixture of beta distributions. Yet, this correction seems to work rather well in practice even in the case of misspecification.

Figure 1 presents a comparison, for the Theil and Duncan index, between the sharp bounds, the naive approach and the corrections proposed by Carrington & Troske (1997), Allen et al. (2009) and Rathelot (2011) when the true data generating process on p is

$$f_p(p) \propto 40p - 83p^2 + 46p^3 - 3p^4 + p^5. \tag{2.3}$$

A first striking point is that the length of the identification region shrinks very quickly with K for the Theil index (less so for the Duncan index). As expected, the naive approach is well above the upper bound of the identification region. In the case of the Theil, the corrected index proposed by Allen et al. (2009) never lies inside the identification interval, while that of Carrington & Troske (1997) (resp. of Rathelot, 2011) is outside for $K \geq 3$ (resp. for $K \geq 5$). The correction proposed by Allen et al. (2009) works better for the Duncan index, being close to the upper bound of the identification interval for most K. On this index, the corrected index of Carrington & Troske (1997), on the other hand, is quite far below the lower bound of this interval, especially for $K \geq 4$. With this particular data generating process, the parametric method of Rathelot, 2011 lies within the bound for all $K \leq 10$. Overall, the approach taken by Rathelot (2011) leads to reasonable corrections when $K \leq 4$, but the parametric misspecification leads to larger bias after. On the contrary, the corrections of Carrington & Troske (1997) and Allen et al. (2009) tend to the true parameter as K becomes large. The convergence is faster for the latter, which is not surprising given its aforementioned properties.

Figure 1: Comparison between the sharp bounds, the naive approach and previous corrections for the Theil and Duncan indices.



Note: the true distribution of p is $f_p(p) \propto 40p - 83p^2 + 46p^3 - 3p^4 + p^5$.

3 Estimation

3.1 Estimation of moments

In this section, we suppose to have in hand an i.i.d. sample $(X_1, ..., X_n)$ for n units. As previously, we also suppose for simplicity that $K_1 = ... = K_n = K$. If K is random, this framework still applies to the subsamples of units sharing the same size. We first address the estimation of the moment vector $m_0 = (E(p^1), ..., E(p^K))$. Equations (2.1) may be written as moment conditions $E\left[h(X, m_0)\right] = 0$, with $h = (h_1, ..., h_K)'$ and, for any

$m = (m_1, ..., m_K)$,

$$h_k(X, m) = \mathbb{1}\{X = k\} - \sum_{j=1}^{K} \binom{K}{j} \binom{j}{k} (-1)^{j-k} m_j.$$

Thus, a first possibility is to estimate m_0 by unconstrained GMM. m_0 is just identified by the K equations in this case, and the GMM estimator $\widetilde{m}$ simply satisfies

$$\widetilde{m} = Q^{-1}\widehat{P},$$

where Q is the $K \times K$ matrix of typical element $\binom{K}{j}\binom{j}{i}(-1)^{j-i}$, $\widehat{P} = (\widehat{P}_1, ..., \widehat{P}_K)'$, with

$$\widehat{P}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = k\}.$$

By the central limit theorem, we directly get

$$\sqrt{n}\left(\widetilde{m} - m_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right), \tag{3.1}$$

where $\Sigma = Q^{-1}RQ^{-1\prime}$ and R is the $K \times K$ matrix of typical element $P_i(\mathbb{1}\{i = j\} - P_j)$.

When K is random, the previous estimator is obtained using only units of same size. It is possible to improve its accuracy by supposing that p is independent of K.[4] In this case indeed, the vector of moments $m_0 = (m_{01}, ..., m_{0\overline{K}})$ (with $\overline{K}$ the maximum of K) is overidentified by $\widetilde{Q}m_0 = \widetilde{P}$, where $\widetilde{P}$ (resp. $\widetilde{Q}$) stacks together vectors $P^K = (\Pr(X = 1|K), ..., \Pr(X = K|K))$ (resp. matrices Q^K of typical element $\binom{K}{j}\binom{j}{i}(-1)^{j-i}$) for different K, and estimate it by minimum distance (see, e.g., Wooldridge, 2002).[5] The subsequent analysis can be conducted similarly, with K replaced by $\overline{K}$ and the variance of $\widetilde{m}$ modified suitably.

The simplicity of $\widetilde{m}$ makes this estimator attractive. However, it suffers from the important drawback of not necessarily belonging to $\mathcal{M}$, even if it will in general with probability approaching one as n grows to infinity.[6] This is all the more problematic for estimating bounds on θ_0 that $\mathcal{P}_{\widetilde{m}}$ is empty in this case. To overcome this issue, we consider afterwards the constrained GMM estimator $\widehat{m}_W$ defined by

$$\widehat{m}_W = \arg\min_{m \in \mathcal{M}} (\widehat{P} - Qm)' W (\widehat{P} - Qm),$$

where W is a definite positive matrix. $\widehat{m}_W$ is well defined as the unique projection onto the closed convex set $\mathcal{M}$. In the absence of ambiguity, we simply let $\widehat{m} = \widehat{m}_W$ hereafter.

[4] Besides, we can test for a weak version of independence between p and K by testing whether subvectors of moments are equal accross different unit sizes.

[5] It is also posible to rewrite the estimating equations as moments, to fit within the GMM framework.

[6] The only exception is when m_0 lies on the boundary of $\mathcal{M}$.

When $\widetilde{m} \in \mathcal{M}$, which occurs with probability approaching one when m_0 is in the interior $\overset{\circ}{\mathcal{M}}$ of $\mathcal{M}$, we have $\widehat{m} = \widetilde{m}$. As a result, the asymptotic properties of $\widehat{m}_W$ are the same as those of $\widetilde{m}$ when $m_0 \in \overset{\circ}{\mathcal{M}}$, for any definite positive matrix W.[7] The advantage of the constrained estimator is that $\mathcal{P}_{\widehat{m}}$ is never empty. Proposition 3.1 makes this point even more precise.

Proposition 3.1 *$\mathcal{P}^k_{\widehat{m}} \neq \emptyset$ for all $k \geq L+1$ (where L is the integer part of $K/2$). Moreover, when $\widehat{m} \neq \widetilde{m}$, $\mathcal{P}^k_{\widehat{m}}$ is reduced to a single distribution with at most $L+1$ points of support, for all $k \geq L+1$.*

The first part of the proposition ensures that provided that $k \geq L+1$, we can always compute the bounds $\widehat{\overline{\theta}^k} = \sup_{F \in \mathcal{P}^k_{\widehat{m}}} g(F)$ or $\widehat{\underline{\theta}^k} = \inf_{F \in \mathcal{P}^k_{\widehat{m}}} g(F)$ that will be our estimators of $\overline{\theta}^k_0$ and $\underline{\theta}^k_0$. The second part of Proposition 3.1, which is based on Wood (1999)'s work on the geometric properties of the binomial mixture problem, shows that when $\widehat{m} \neq \widetilde{m}$, these estimated bounds are actually equal, so that the estimated identification interval is reduced to a singleton.

To compute $\widehat{m}$ in practice, we proceed in two steps. First we check whether $\widetilde{m} \in \mathcal{M}$, using Proposition 2.1 above. If this is the case, we have of course $\widehat{m} = \widetilde{m}$. If not, we might use the same proposition to obtain $\widehat{m}$, but this is computationally difficult, as it involves a nonlinear optimization with nonlinear inequality constraints. We rather rely on Proposition 3.1, which implies that there is a unique distribution, with at most $L+1$ support points, that corresponds to $\widehat{m}$. Such a distribution can be described by an ordered vector $x^* = (x^*_1, ..., x^*_{L+1})' \in \mathcal{S}_{L+1} = \{(x_1, ..., x_{L+1}) : 0 \leq x_1 < ... < x_{L+1} \leq 1\}$ and a vector of corresponding probabilities $y^* = (y^*_1, ..., y^*_{L+1})' \in \mathcal{T}_{L+1} = \{(y_1, ..., y_{L+1}) \in [0,1]^{L+1} : \sum_{k=1}^{L+1} y_k = 1\}$. For all $x \in \mathcal{S}_{L+1}$, let

$$B(x) = \begin{pmatrix} x_1 & \dots & x_{L+1} \\ & \vdots & \\ x_1^K & \dots & x_{L+1}^K \end{pmatrix},$$

so that the vector of moment of $(x,y) \in \mathcal{S}_{L+1} \times \mathcal{T}_{L+1}$, is $B(x)y$. Thus, by definition of $\widehat{m}$ and Proposition 3.1, (x^*, y^*) can be obtained by

$$(x^*, y^*) = \arg \min_{(x,y) \in \mathcal{S}_{L+1} \times \mathcal{T}_{L+1}} (\widehat{P} - QB(x)y)'W(\widehat{P} - QB(x)y), \qquad (3.2)$$

[7] Hence, the choice of W has no impact on the asymptotic distribution of $\widehat{m}_W$ unless m_0 is at the boundary of $\mathcal{M}$. On the other hand, this choice matters in the test of the model, see Subsection 2.3 below.

and $\widehat{m} = B(x^*)y^*$.[8] As a result, the computation of $\widehat{m}_W$ involves an optimization over a quite low dimensional space $(2(L+1))$ under linear equality and inequality constraints only.[9] It turns out to be quick to compute in practice.

3.2 Estimation of the bounds on θ_0

We now turn to estimation of the bounds θ_0, for any $k \geq 1$. We first study estimators of $\overline{\theta}_0^k$ and $\underline{\theta}_0^k$. This will allow us to yield asymptotically exact confidence intervals on θ_0 when g is linear because, by Theorem 2.2, $\overline{\theta}_0^{K+1} = \overline{\theta}_0$ and $\underline{\theta}_0^{K+1} = \underline{\theta}_0$.

Any $F \in \mathcal{P}_m^k$ is defined by its support points $x = (x_1, ..., x_k) \in \mathcal{S}_k$ and associated probabilities $y = (y_1, ..., y_k)' \in \mathcal{T}_k$. For a given vector of moments m, the moment constraints write $A(x)y = (1, m')'$, where $A(x)$ is the matrix

$$A(x) = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_k \\ & \vdots & \\ x_1^K & \dots & x_k^K \end{pmatrix}.$$

When $F \in \mathcal{P}_m^k$, we may rewrite g as a function of x, m and y, which we denote $q^k(x, y, m)$.[10] The bounds on θ then satisfy

$$\overline{\theta}^k(m) = \max_{(x,y)\in\mathcal{S}_k\times[0,1]^k} q^k(x,y,m) \quad \text{s. t. } A(x)y = (1,m')', \tag{3.3}$$

$$\underline{\theta}^k(m) = \min_{(x,y)\in\mathcal{S}_k\times[0,1]^k} q^k(x,y,m) \quad \text{s. t. } A(x)y = (1,m')'. \tag{3.4}$$

We simply define our estimator of $\overline{\theta}_0^k = \overline{\theta}^k(m_0)$ and $\underline{\theta}_0^k$ by

$$\widehat{\overline{\theta}}^k = \overline{\theta}^k(\widehat{m}), \ \widehat{\underline{\theta}}^k = \underline{\theta}^k(\widehat{m}).$$

As mentioned before, when $\widehat{m} \neq \widetilde{m}$, there is a unique distribution in $\mathcal{P}_m^k$ which rationalizes the vector $\widehat{m}$. In this case, using the same notations as previously, we have $\widehat{\underline{\theta}}^k = \widehat{\overline{\theta}}^k = q^k(x^*, y^*, \widehat{m})$, and no optimization is needed.

Our asymptotic result is based on the following regularity conditions.

[8] When the distribution that rationalizes $\widehat{m}$ has less than $L+1$ support points, Program (3.2) does not admit a unique solution because we can set some components of y to zero and move freely the corresponding components of x. In this case any solution can be chosen, since it leads to the same $\widehat{m}$ anyway.

[9] In practice, we cannot use the strict inequalities in $\mathcal{S}_{L+1}$ in our optimization. We approximate this set by $\{(x_1, ..., x_{L+1}) : 0 \leq x_1 \leq x_2 - \varepsilon ... < x_{L+1} - L\varepsilon \leq 1 - L\varepsilon\}$, for a small enough constant $\varepsilon > 0$.

[10] g may depend directly on m and not only through x and y. For instance the Theil index may be written as $1 - y'(x\ln(x))/(m_1 \ln(m_1))$ (where, for any $x \in \mathbb{R}^k$, $x\ln(x)$ is defined componentwise).

Assumption 3.1 *The global maximum $(\overline{x}_0, \overline{y}_0)$ corresponding to (3.3) for $m = m_0$ is such that q^k is C^2 in a neighborhood of $(\overline{x}_0, \overline{y}_0, m_0)$, the subgradient of binding constraints taken at $(\overline{x}_0, \overline{y}_0)$ is full rank and for all $u \neq 0$ such that $\frac{\partial}{\partial(x,y)}\left[A(x)y\right]_{|(\overline{x}_0,\overline{y}_0)}.u = 0$, we have $u'Hu < 0$, H being the Hessian of the Lagrangian corresponding to Program (3.3), taken at $(\overline{x}_0, \overline{y}_0)$. The same holds for the global minimum $(\underline{x}_0, \underline{y}_0))$ corresponding to (3.4).*

Assumption (3.1) ensures, by the smooth maximum theorem (see Carter, 2001), that $\overline{\theta}^k$ and $\underline{\theta}^k$ are C^1 in a neighborhood of m_0. Moreover, by the envelope theorem, the derivatives of $\overline{\theta}^k$ and $\underline{\theta}^k$ take a simple form. Then, by the delta method and asymptotic normality of $\widehat{m}$, $(\overline{\theta}^k, \underline{\theta}^k)$ is also asymptotically normal, and its asymptotic variance has a simple expression, which can be consistently estimated. In the following, we let $\widehat{\Sigma} = Q^{-1}\widehat{R}Q^{-1}$, $\widehat{R}$ being the matrix whose (i,j) element equals $\widehat{P}_i(\mathbb{1}\{i = j\} - \widehat{P}_j)$.

Theorem 3.2 *Suppose that $m_0 \in \overset{\circ}{\mathcal{M}}$ and Assumption 3.1 holds. Then*

$$\sqrt{n}\begin{pmatrix} \widehat{\overline{\theta}}^k - \overline{\theta}_0^k \\ \widehat{\underline{\theta}}^k - \underline{\theta}_0^k \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0,\ J_k \Sigma J_k'\right)$$

with

$$J_k = \begin{pmatrix} \frac{\partial \overline{\theta}^k}{\partial m}(m_0) \\ \frac{\partial \underline{\theta}^k}{\partial m}(m_0) \end{pmatrix} = \begin{pmatrix} \frac{\partial q^k}{\partial m}(\overline{x}_0, \overline{y}_0, m_0) - \overline{\lambda}_0^{k}{}' \\ \frac{\partial q^k}{\partial m}(\underline{x}_0, \underline{y}_0, m_0) - \underline{\lambda}_0^{k}{}' \end{pmatrix},$$

and $\overline{\lambda}_0^k$ and $\underline{\lambda}_0^k$ are the vector of Lagrange multiplier corresponding to the first order conditions of (3.3) and (3.4). Moreover, letting $\widehat{J}_k = \begin{pmatrix} \frac{\partial \overline{\theta}^k}{\partial m}(\widehat{m}) \\ \frac{\partial \underline{\theta}^k}{\partial m}(\widehat{m}) \end{pmatrix}$, we have

$$\widehat{J}_k \widehat{\Sigma} \widehat{J}_k \xrightarrow{P} J_k \Sigma J_k'.$$

Theorem 3.2 can be applied for instance to the Theil or Gini index, for which it is easy to see that the regularity conditions on q^k hold. On the other hand, the smoothness condition is not satisfied for the lower bound of the Duncan index. Indeed, m_{01} belongs in general to the support of $\underline{x}_0$, making q^k not C^2 in the neighborhood of $(\underline{x}_0, m_0)$. However, we can still prove that $\underline{\theta}(.)$ is C^1. The idea is that for any value of m near m_0, m will be one of the support points of the distribution that solves (3.4). We may thus rewrite (3.4) with one dimension less on x. Moreover, q^k is C^2 in the neighborhood of the new solution, and we can then apply the smooth maximum theorem.

When g is linear, we can combine Theorem 3.2 with Theorem 2.2 to yield asymptotically exact confidence intervals on θ_0. We let z_α denote the quantile of order α of a standard normal variable and define

$$\mathrm{CI}^1_{1-\alpha} = \left[\widehat{\underline{\theta}}^{K+1} - z_\alpha \frac{\partial \underline{\theta}^{K+1}}{\partial m'}(\widehat{m})\widehat{\Sigma}\frac{\partial \underline{\theta}^{K+1}}{\partial m}(\widehat{m}),\ \widehat{\overline{\theta}}^{K+1} + z_\alpha \frac{\partial \overline{\theta}^{K+1}}{\partial m'}(\widehat{m})\widehat{\Sigma}\frac{\partial \overline{\theta}^{K+1}}{\partial m}(\widehat{m})\right], \quad (3.5)$$

Corollary 3.3 *Suppose that g is linear, $m_0 \in \overset{\circ}{\mathcal{M}}$, $\underline{\theta}_0 < \overline{\theta}_0$ and Assumption 3.1 holds. Then*

$$\inf_{\theta_0 \in [\underline{\theta}_0, \overline{\theta}_0]} \lim_{n \to \infty} \Pr(\theta_0 \in CI^1_{1-\alpha}) = 1 - \alpha.$$

As noted by Imbens & Manski (2004), such intervals do not provide a correct asymptotic level uniformly over all possible distributions because, intuitively, its asymptotic level is only $1 - 2\alpha$ when $\underline{\theta}_0 = \overline{\theta}_0$. Yet, we cannot apply Imbens & Manski's improved confidence interval here because asymptotic normality fails to hold uniformly. Indeed, $\widehat{m}$ is not asymptotically normal in general when m_0 lies on the boundary of $\mathcal{M}$. It is however possible to define another confidence interval which is valid under less restrictive conditions. For that purpose, let $I_{1-\alpha}$ denote a confidence region on m_0 with asymptotic level $1-\alpha$. A natural one is

$$I_{1-\alpha} = \{m \in \mathcal{M} : \|m - \widetilde{m}\| \leq \chi^2_K(1-\alpha)\},$$

where $\|x\| = x'\widehat{\Sigma}^{-1}x$ and $\chi^2_K(1-\alpha)$ is the $1-\alpha$ quantile of a χ^2_K distribution. Then define

$$\mathrm{CI}^2_{1-\alpha} = \left[\inf_{m \in I_{1-\alpha}} \underline{\theta}^{K+1}(m), \sup_{m \in I_{1-\alpha}} \overline{\theta}^{K+1}(m) \right]. \tag{3.6}$$

Proposition 3.4 *Suppose that g is linear. Then*

$$\lim_{n \to \infty} \Pr([\underline{\theta}_0, \overline{\theta}_0] \subset CI^2_{1-\alpha}) \geq 1 - \alpha.$$

The validity of $\mathrm{CI}^2_{1-\alpha}$ is obtained under very mild assumptions. In particular, no regularity condition is required on q^k. Even if these conditions hold, another advantage of $\mathrm{CI}^2_{1-\alpha}$ is that it does not require to estimate $\partial\overline{\theta}^{K+1}/\partial m'(m_0)$, contrary to $\mathrm{CI}^1_{1-\alpha}$. This is especially convenient when $\widehat{m}$ lies on the boundary of $\mathcal{M}$ (i.e., when $\widetilde{m} \notin \mathcal{M}$), because in this case, $\partial\overline{\theta}^{K+1}/\partial m'(\widehat{m})$ may not exist.[11] As shown in the Monte Carlo simulations presented in the next section, this happens very often for typical sample sizes when $K \geq 8$. The drawback of these confidence intervals is that they are conservative in general, even though the simulations suggest that they may actually be still very informative.

The previous confidence intervals may also be used for testing issues. In our application below, we are interested in particular by the equality of the index over a range of K. Given $K_1 \dots K_v$, consider the null hypothesis, consisting of the equality of v terms:

$$\theta_0(K_1) = \theta_0(K_2) = \cdots = \theta_0(K_v),$$

[11] Indeed, if $\widehat{m}$ lies on the boundary of $\mathcal{M}$, $\overline{\theta}^{K+1}(\widehat{m} + c\delta)$ may not exist for some directions δ, for any $c \in \mathbb{R}$.

where $\theta_0(K)$ is the true parameter corresponding to units of size K. We base our tests on the intersection of the confidence intervals for these parameters. Let θ_0 denote $\theta_0(K), K = K_1, ..., K_v$, under the null and let $\mathrm{CI}_{(1-\alpha)^{1/v}}(K)$ denote an asymptotically conservative confidence region at the level $(1-\alpha)^{1/v}$ for $\theta_0(K)$, we get under the null, by independence of the confidence intervals between the different subsamples,

$$\begin{aligned}
\Pr\left(\cap_{k=1}^{v}\mathrm{CI}_{(1-\alpha)^{1}/v}(K_k) \neq \emptyset\right) &= \Pr\left(\exists\theta : \theta \in \cap_{k=1}^{v}\mathrm{CI}_{(1-\alpha)^{1/v}}(K_k)\right) \\
&\geq \Pr\left(\theta_0 \in \cap_{k=1}^{v}\mathrm{CI}_{(1-\alpha)^{1/v}}(K_k)\right) \\
&\geq \prod_{k=1}^{v}\Pr\left(\theta_0 \in \mathrm{CI}_{(1-\alpha)^{1/v}}(K_k)\right),
\end{aligned}$$

so that the probability that the intersection is non-empty is asymptotically greater than $1-\alpha$ under the null. It is also easy to see that this test is consistent against alternatives where $\cap_{k=1}^{v}\left[\underline{\theta}_0(K_k), \overline{\theta}_0(K_k)\right] = \emptyset$.

Finally, in the case of nonlinear functionals, we may also use $\mathrm{CI}^1_{1-\alpha}$ or $\mathrm{CI}^2_{1-\alpha}$, with a $k \neq K+1$. A difference, however, is that the asymptotic level of such confidence intervals is lower than α. This is because $\overline{\theta}_0^k < \overline{\theta}_0$ and $\underline{\theta}_0^k > \underline{\theta}_0$ in general. A solution would be to make k tend to infinity with n at a sufficient rate, so that $\overline{\theta}_0 - \overline{\theta}_0^k$ would become negligible with respect to the variance of $\widehat{\overline{\theta}}^k$. This issue is left for future research.

3.3 Test of the binomial mixture model

As mentioned previously, the binomial model is testable. In this subsection, we develop a simple test of this hypothesis that, as shown before, is equivalent to $m_0 \in \mathcal{M}$. The idea is to approximate the distance between m_0 and $\mathcal{M}$ by the one between $\widetilde{m}$ and $\mathcal{M}$. This leads to a consistent test because $\widetilde{m}$ estimates consistently m_0 in both the null and alternative hypothesis. Formally, we have

$$\sqrt{n}\left(\widetilde{m} - m_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right).$$

Thus $n\left\|\widetilde{m} - m_0\right\| \to \chi^2_K$. Besides, under the null hypothesis that $m_0 \in \mathcal{M}$, $\left\|\widetilde{m} - m_0\right\| \geq \left\|\widetilde{m} - \widehat{m}_{\widehat{R}^{-1}}\right\|$ by definition of $\widehat{m}_{\widehat{R}^{-1}}$. Thus, letting $S_n = n\left\|\widetilde{m} - \widehat{m}_{\widehat{R}^{-1}}\right\|$, we have, under the null hypothesis,

$$\sup_{m_0 \in \mathcal{M}}\ \limsup_{n\to\infty}\Pr\left(S_n > \chi^2_K(1-\alpha)\right) \leq \alpha.$$

Besides, by the triangular inequality,

$$\left\|\widetilde{m} - \widehat{m}_{\widehat{R}^{-1}}\right\| \geq \left\|m_0 - \widehat{m}_{\widehat{R}^{-1}}\right\| - \left\|m_0 - \widetilde{m}\right\|.$$

Under the alternative, the first term on the right-hand side converges by the continuous mapping theorem to a strictly positive number, which is the distance between m_0 and $\mathcal{M}$. Thus, $S_n \to \infty$ under the alternative, and the test is consistent.

An issue with this test is that it is conservative in general. To improve its level, a possibility may be to rely on bootstrap, using the following procedure:

1. Estimate $\widetilde{m}$ and $\widehat{m}_{\widehat{R}^{-1}}$. Compute S_n and $\widehat{P}_2 = Q\widehat{m}_{\widehat{R}^{-1}}$.

2. Compute the distribution of S_n^* that we obtain when $P = (\Pr(X = 1), ..., \Pr(X = K))$ is equal to $\widehat{P}_2$. For that purpose, we can simulate iid samples of size n of X_i^* satisfying $\Pr(X_i^* = k) = \widehat{P}_{2k}$, $k \in \{1, ..., K\}$ and $\Pr(X_i^* = 0) = 1 - \sum_{k=1}^K \widehat{P}_{2k}$.

3. Compare S_n with the $(1 - \alpha)$th quantile of S_n^*.

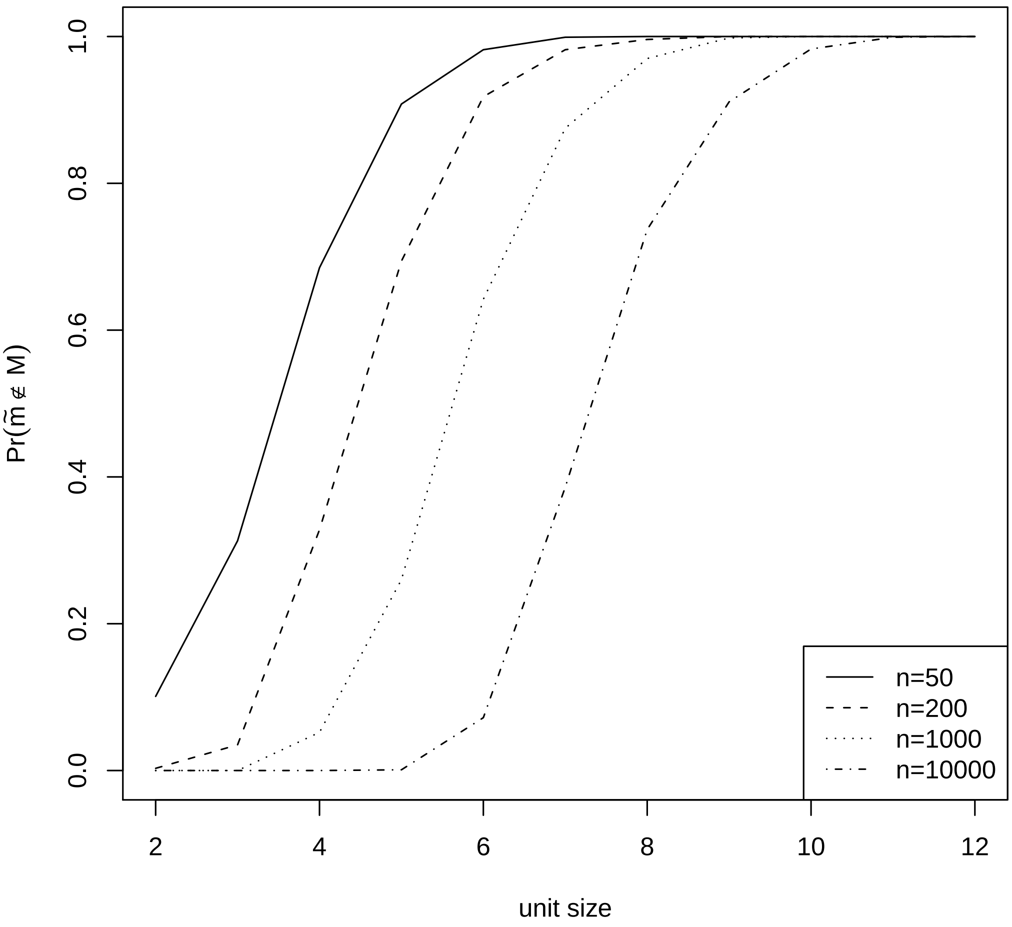
The idea of this procedure is to compute the distribution of S_n under the null (because $\widehat{m}_{\widehat{R}^{-1}} \in \mathcal{M}$) and if the true vector P is equal to $\widehat{P}_2$. Because $\widehat{P}_2$ converges to P under the null, the bootstrap test will have asymptotically the nominal level if the distribution of S_n is continuous with respect to P. Even if we do not address this issue here, our simulations below are encouraging, the true level being close to the nominal one. We thus also rely on this test in our application in Section 5.

4 Monte Carlo simulations

This section presents the results of Monte Carlo simulations designed to assess the performance of the method presented in this paper in order to solve small-unit biases.

We first study whether the constraint that m_0 belongs to $\mathcal{M}$ is binding in practice when estimating m_0. The data generating process is defined by (2.3), and we estimate $\Pr(\widetilde{m} \in \mathcal{M})$ for different sample and unit sizes. Figure 2 presents the results for $n \in \{50, 200, 1\,000, 10\,000\}$ and $K \in \{2, ..., 12\}$. For any n, the probability grows quite quickly to one with K. This reflects the aforementioned fact that the set $\mathcal{M}$ shrinks very quickly with K. For instance, with 200 units, the estimated probability (with 1,000 simulations) is one as soon as K is equal to 7. When the sample contains more units, the probability is systematically lower because the estimation precision increases, but for $K \geq 10$, this probability remains very close to 1 for samples as large as 10,000. This implies that for $K \geq 10$, we should expect to generally get a point estimate of the identification region of θ_0, even though this true identification region is not reduced to a singleton.[12] An interpretation of this is that the length of the true identification interval for such values of K and n is far below the length due to estimation. Our ignorance on the true parameter mostly stems from finite sampling rather than partial identification issues.

Figure 2: Probability that the estimated moments $\tilde{m} \notin \mathcal{M}$.



Note: each dot corresponds to 1,000 simulations drawn with the distribution defined in (2.3).

Tables 1 and 2 display the properties of $[\underline{\widehat{\theta}}, \widehat{\overline{\theta}}]$ and the confidence intervals $\mathrm{CI}^1_{1-\alpha}$ and $\mathrm{CI}^2_{1-\alpha}$ for different n and K, for the Theil index and still for the data generating process defined

[12] This result is in line with the Monte Carlo simulations of Wood (1999), who focuses on the distribution of p and estimates it with either a projection method or maximum likelihood. As here, his estimator is unique when $\widetilde{m} \notin \mathcal{M}$. He shows that this holds generally for moderate to large K, even if n is large.

by (2.3). For this distribution, $\theta_0 \simeq 0.1654$. $\mathrm{CR}(\theta_0)$ is the coverage rate of the true parameter by the confidence intervals, while $\mathrm{CR}([\underline{\theta}_0, \overline{\theta}_0])$ is the coverage rate of the whole identification interval. We do not present $\mathrm{CI}^1_{1-\alpha}$ for $K \geq 6$ as very often for such values of K, $\widetilde{m} \notin \mathcal{M}$. As explained above, this causes trouble in the computation of this confidence interval. Overall, the estimator of the identification interval is quite precise even for small samples. In our setting, we only observe a significant bias on $\overline{\theta}_0$, which however does not lead to a low coverage of the confidence intervals. Consistent with Figures 1 and 2, we see that even for $n = 10,000$, standard errors are far larger than the length of the identification region for $K \geq 9$. This means that for $K \geq 9$, uncertainty mostly stems from estimation, not from partial identification.

Table 1: Performance of the estimator of $[\underline{\theta}_0, \overline{\theta}_0]$ and coverage rates of $\mathrm{CI}^2_{1-\alpha}$.

				$\mathrm{CI}^2_{1-\alpha}$		
K	$[\underline{\theta}_0, \overline{\theta}_0]$	n	$[E(\underline{\theta}_0), E(\overline{\theta}_0)]$ $(\sigma(\underline{\theta}_0))$ $(\sigma(\overline{\theta}_0))$	Length	$\mathrm{CR}(\theta_0)$	$\mathrm{CR}([\underline{\theta}_0, \overline{\theta}_0])$
3	$[0.154; 0.212]$	100	$[0.162; 0.204]$ (0.06) (0.077)	0.400	1	0.995
		1,000	$[0.155; 0.212]$ (0.016) (0.023)	0.180	1	1
		10,000	$[0.155; 0.213]$ (0.006) (0.009)	0.098	1	0.985
6	$[0.164; 0.170]$	100	$[0.162; 0.162]$ (0.047) (0.047)	0.273	1	1
		1,000	$[0.169; 0.170]$ (0.013) (0.014)	0.098	1	1
		10,000	$[0.165; 0.169]$ (0.004) (0.005)	0.037	1	0.995
9	$[0.165; 0.166]$	100	$[0.166; 0.166]$ (0.038) (0.038)	0.234	1	1
		1,000	$[0.165; 0.165]$ (0.013) (0.013)	0.081	1	1
		10,000	$[0.166; 0.166]$ (0.004) (0.004)	0.028	1	1
12	$[0.165; 0.166]$	100	$[0.162; 0.162]$ (0.032) (0.032)	0.221	1	1
		1,000	$[0.166; 0.166]$ (0.011) (0.011)	0.076	1	1
		10,000	$[0.165; 0.165]$ (0.004) (0.004)	0.025	1	1

Note: for each distribution, simulations are based on 200 draws of samples. The true distribution of p is $f_p(p) \propto 40p - 83p^2 + 46p^3 - 3p^4 + p^5$, leading to $\theta_0 \simeq 0.1654$.

As expected, the confidence interval $\mathrm{CI}^2_{1-\alpha}$ is conservative when looking at the coverage rate of either the true parameter or the true identification interval. Less expectedly, we observe a similar pattern for $\mathrm{CI}^1_{1-\alpha}$ for small sample sizes. Not surprisingly, $\mathrm{CI}^1_{1-\alpha}$ is smaller on average than $\mathrm{CI}^2_{1-\alpha}$. But the difference between the two is not that large. Because $\mathrm{CI}^2_{1-\alpha}$ can be computed whether $m_0 \in \mathcal{M}$ or not, this suggests to use this confidence interval in practice.

Table 2: Comparison between $\mathrm{CI}^1_{1-\alpha}$ and $\mathrm{CI}^2_{1-\alpha}$ for $K = 3$.

	$\mathrm{CI}^1_{1-\alpha}$			$\mathrm{CI}^2_{1-\alpha}$		
n	Length	$\mathrm{CR}(\theta_0)$	$\mathrm{CR}([\underline{\theta}_0, \overline{\theta}_0])$	Length	$\mathrm{CR}(\theta_0)$	$\mathrm{CR}([\underline{\theta}_0, \overline{\theta}_0])$
100	0.312	0.985	0.975	0.400	1	0.995
1,000	0.131	0.995	0.975	0.180	1	1
10,000	0.081	1	0.84	0.098	1	0.985

Note: for each distribution, simulations are based on 200 draws of samples. The true distribution of p is $f_p(p) \propto 40p - 83p^2 + 46p^3 - 3p^4 + p^5$, leading to $\theta_0 \simeq 0.1654$.

Finally, Table 3 displays some elements about the performance of the conservative and the bootstrap test of the binomial model proposed in the previous section. We use a data generating process such that m_0 is at the boundary of $\mathcal{M}$, namely a discrete distribution for p with values $0.25, 0.5$ and 0.75 and corresponding probabilities 0.4, 0.2 and 0.4. As expected, the conservative test exhibits levels which are quite below the nominal one. On the contrary, the bootstrap test seems to perform well here, with true levels close to the nominal one in general.

Table 3: Tests of the binomial model: true levels of the conservative and bootstrap test (for a nominal level of 5%).

K	n	conservative	bootstrap
6	100	0	0,05
	1,000	0	0,045
	10,000	0	0,035
9	100	0,01	0,055
	1,000	0	0,04
	10,000	0	0,055
12	100	0,02	0,035
	1,000	0,005	0,065
	10,000	0	0,06
15	100	0,01	0,035
	1,000	0,01	0,055
	10,000	0,01	0,09

Note: For each distribution, simulations are based on 200 draws of samples. The true distribution of p takes values $0.25, 0.5$ and 0.75 with probability 0.4, 0.2 and 0.4 respectively.

5 An application to workplace segregation by nationality across French establishments

Understanding why and how employers make their hiring decisions and employees apply for jobs implies to be able to measure workplace segregation. For example, the issue of segregation is also related to employment and wage differentials across groups, either on sex or ethnic grounds. Early works focused on gender or race segregation across occupations or industries, see, *e.g.*, Fields & Wolff (1991). Groshen (1991) is the first contribution to use the information available at the scale of establishments. Carrington & Troske (1995) use the 1983 CPS to compute Duncan indices for gender segregation across establishments, with a focus on small firms. Another strand of literature, which aims at linking skill dispersion with wage distribution, requires the computation of segregation indices. Kremer & Maskin (1996) and Kramarz et al. (1996) analyze, in the US and the French cases, how skill dispersion, measured with segregation indices, accounts for changes in the wage structure. Iranzo et al. (2008) investigate a similar issue in the case of Italy and find that most of overall skill dispersion is within, not between, firms.

However, few of these works acknowledge the issue of small-unit bias and them attempt to correct the indices.[13] Carrington & Troske (1997) present new results on black/white segregation introducing a method to correct for small-bias unit. Hellerstein & Neumark (2008) use the 1990 Decennial Employer-Employee Database to measure workplace segregation by education, language and ethnicity. They compute adjusted indices using Carrington and Troske's method.

In this section, we aim to compute a Theil index to measure the segregation between French and foreigners across French businesses. Do all establishments have the same share of foreigners or, on the contrary, do some firms specialize in hiring foreign workers while the other ones avoid those? As a large share of workers are employed in small establishments, not taking into account the small unit bias would certainly lead to upward-biased estimates of segregation levels. We use the method introduced in this paper to compute either point or set estimate of the Theil index. 95% confidence intervals were computed using CI^1 whenever $\widetilde{m} \in \mathcal{M}$ and CI^2 otherwise. As a matter of comparison, we also display the naive estimate and the ones proposed by Carrington & Troske (1997), Allen et al. (2009) and Rathelot (2011).

We use the *Déclaration Annuelles de Données Sociales* (DADS) for year 2007, the French matched employer-employee database, which is exhaustive on the private sector (1.8 million establishments). We restrict ourselves to the 1.65 million establishments with less than 25 employees. Two minority groups are considered successively: individuals born outside France with the citizenship of a country belonging to the European Union, and individuals born outside France with the citizenship of a country which does not belong to the European Union. The first group will be called Europeans to simplify, and the other will be called non-Europeans. In France, the most important groups of migrants come from Southern Europe and North Africa (Insee, 2005). More specifically, in 1999, migrants from Italy, Portugal and Spain amounted to 30% of all migrants. They also represent 66% of the European migrants: our European sample is thus likely to be dominated by Southern Europeans. Migrants from Algeria, Morocco and Tunisia amount to another 30% of all migrants while 10% come from other African countries and 13% from Asia. Therefore, Africa represent more than 70% of non-European immigrant origins.

[13] Kremer & Maskin (1996) and Kramarz et al. (1996) interpret their segregation measure as a R-squared and suggest that using adjusted R-squared might be a way to deal with small-unit issues.

Table 4: Test of the binomial mixture model for $K \geq 8$.

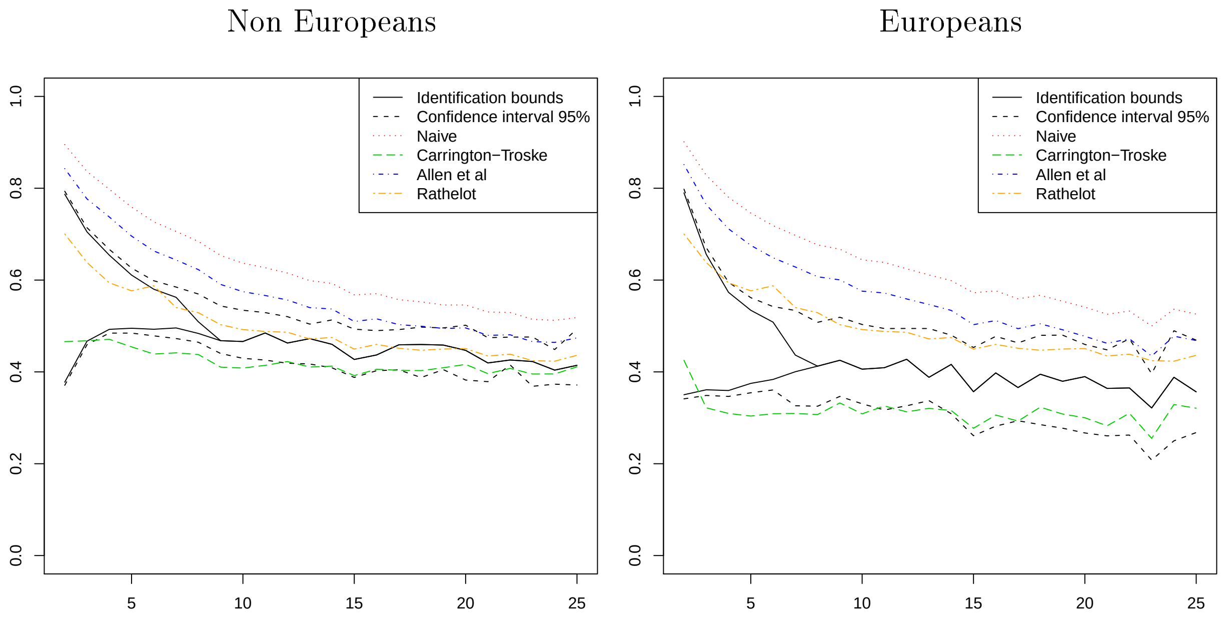
Size	P-value of the bootstrap test	
K	Non-European	European
8	1	0.77
9	0.12	0.79
10	0.27	0.75
11	0.34	0.98
12	0.44	0.26
13	0.03	0.01
14	0.54	0.37
15	0	0.78
16	0.19	0.91
17	0.22	0.17
18	0.6	0.33
19	0.43	0.6
20	0.97	0.37
21	0.48	0
22	0.91	0.06
23	0.07	0.44
24	0.26	0.55
25	0.95	0.24

Before presenting our results, we first check that the binomial mixture model is not rejected in these data. For $K = 2...7$, we obtain that $\widetilde{m} \in \mathcal{M}$ in both samples, so the test is automatically accepted. For $K \geq 8$, on the other hand, $\widetilde{m} \notin \mathcal{M}$. Of course, one should not conclude that the binomial mixture model should be discarded. The Monte Carlo simulations presented in the previous section emphasize that $\Pr(\widetilde{m} \notin \mathcal{M})$ is very large for $K \geq 9$ even when this model is true and n is as large as $10,000$ (in our application, n lies between 9,000 and 36,700 for $K \geq 8$). We perform the bootstrap test detailed above for $K \geq 8$ (see Table 4). Actually, our test leads to a rejection of the binomial mixture model at the 5% level for only $K = 13$ and $K = 15$ on Non-Europeans, and for $K = 13$ and $K = 21$ for Europeans. We see this as an evidence that the binomial mixture model is reasonable here.

Figure 3 displays the estimates of workplace segregation, using the Theil index, across French establishments for non-Europeans and Europeans. In line with Figure 1, we observe that the sharp bounds become very informative for $K \geq 5$. The estimated identification region reduces to a singleton for $K \geq 8$, as expected since for these values $\widetilde{m} \notin \mathcal{M}$. As in Figure 1, we also observe that the naive estimator, Carrington and Troske's correction and

Allen et al. estimator lie outside our confidence intervals for most values of K and for both samples. Only the method proposed by Rathelot (2011) seems to perform well in practice, suggesting that the mixture of two beta distributions is a reasonable approximation for the distribution of p here.

Figure 3: Theil indices for Non-Europeans and Europeans, by firm size.



Non Europeans

Europeans

The estimated identification regions displayed in Figure 3 show that the segregation level for the non-European sample is 5 to 10 points higher than for the European sample. Interestingly, this pattern does not appear when using either the naive or the Allen et al. estimate. This shows that the bias of such estimators may be sensitive not only to K, but also to the underlying distribution. As a result, using such estimators for comparing segregation between different groups may be misleading.

A striking difference between the naive and Allen et al. estimates, on the one hand, and the identification region we estimate, on the other hand, is that segregation seems to be strongly negatively correlated with K in the first case, very less so in the second case. That the former decrease with K is not surprising, given that their bias directly depends on it (proportional to $1/K$ for the naive estimator, $1/K^{3/2}$ or $1/K^2$ for Allen et al. estimator). But there may still exist a true negative dependence of the segregation level on firm sizes. Small firms may rely more heavily on social networks in their hiring process, for instance, resulting in a higher segregation between firms (people from minority tending to hire other people from the same minority, and conversely).[14] To test for this, we consider the null

[14] Pistaferri (1999) shows that, in Italy, smaller firms tended to use more often informal hiring channels.

hypothesis

$$\theta_0(5) = \theta_0(10) = \theta_0(15) = \theta_0(20) = \theta_0(25),$$

where $\theta_0(K)$ is the true parameter corresponding to firms of size K. For both minority populations, we cannot reject the null hypothesis at the level of 10%. Thus, contrary to what is suggested by the naive and Allen et al. estimates, we cannot reject the hypothesis that segregation levels do not depend on firm sizes.

6 Conclusion

In this paper, we investigate what can be learned on the feature of a random variable p when only an imperfect measure of it, distributed according to a binomial variable $B(K, p)$, is available. We show that in general this leads to partial identification of these features. We then develop inference on the bounds, as well as a test of the binomial model.

Starting there, an interesting avenue of research would be to study the dependence of segregation indices on unit characteristics (such as sectors or geographical areas for firms). We also believe that our method and results have potential interest beyond the precise application considered here. The idea of approximating all distributions by discrete ones only to compute bounds, also pushed forwards by Chernozhukov et al. (2009) in the case of fixed effect nonlinear panel models, is important in practice. We fully justified it here for linear functionals only, so it seems desirable to develop a general methodology that would also include nonlinear functionals.

Additionally, Dustmann et al. (2010) for Germany or Aslund & Skans (2010) for Sweden show that firms are more likely to hire minority workers from a particular group if the existing share of workers from that group employed in the firm is higher. In a similar vein, Giuliano et al. (2009) shows, for the US, that manager race affects the racial composition of new hires.

7 Appendix: proofs

Proof of Theorem 2.2

We show below Part 1 and Part 2. Part 3 follows by applying Part 2 to g and $-g$.

1. Approximation by discrete distributions.

We first prove that $\cup_{k=1}^{\infty}\mathcal{P}_{m_0}^k$ is dense in $\mathcal{P}_{m_0}$. Fix $F \in \mathcal{P}_{m_0}$ and $\varepsilon > 0$, and let us prove that there exists $G \in \cup_{k=1}^{\infty}\mathcal{P}_{m_0}^k$ such that $||G-F||_\infty = \sup_{x\in[0,1]}|G(x)-F(x)| < \varepsilon$. If F is discrete then this is obvious. Otherwise, suppose first that F has no jump larger than ε. Then there exists $a_0 = 0 < a_1 < ... < a_J < a_{J+1} = 1$ such that $F(a_{j+1}) - F(a_j) < \varepsilon$. For all $j = 0, ..., J$, let

$$F_j(x) = \mathbb{1}\{x > a_{j+1}\} + \frac{F(x)-F(a_j)}{F(a_{j+1})-F(a_j)}\mathbb{1}\{x \in [a_j, a_{j+1}]\},$$

and let $m^j = \left(\int xdF_j(x), ..., \int x^K dF_j(x)\right)'$. There exists (see, e.g., Curto & Fialkow, 1991, Theorems 4.1 and 4.3) a discrete distribution with support $[a_j, a_{j+1}]$ such that its c.d.f. G_j satisfies

$$\left(\int xdG_j(x), ..., \int x^K dG_j(x)\right)' = m^j.$$

Then let $G = \sum_{j=0}^J (F(a_{j+1}) - F(a_j))G_j$. By construction, $G \in \cup_{k=1}^{\infty}\mathcal{P}_{m_0}^k$. Moreover, for any $x \in [a_j, a_{j+1})$, $G(x) = (F(a_{j+1}) - F(a_j))G_j(x)$ and $F(x) = (F(a_{j+1}) - F(a_j))F_j(x)$. Thus,

$$|G(x) - F(x)| = (F(a_{j+1}) - F(a_j))|G_j(x) - F_j(x)| < \varepsilon.$$

As a result, $||G-F||_\infty < \varepsilon$. Now, F may admit jumps larger than ε. Let $(x_1, ..., x_l)$ (resp. $(y_1, ..., y_l)$) denote the corresponding points (resp. probabilities). Define the function $\widetilde{F}$ by

$$\widetilde{F}(x) = \frac{F(x) - \sum_{j=1}^l y_j \mathbb{1}\{x \geq x_j\}}{1 - \sum_{j=1}^l y_j}.$$

Because F is not discrete, $\sum_{j=1}^l y_j < 1$ so that $\widetilde{F}$ is well defined. Besides, by construction $\widetilde{F}$ has no jump larger than ε. As a result, there exists $\widetilde{G} \in \cup_{k=1}^{\infty}\mathcal{P}_{\widetilde{m}_0}^k$, where $\widetilde{m}_0 = \left(\int xd\widetilde{F}(x), ..., \int x^K d\widetilde{F}(x)\right)'$ such that $||\widetilde{G} - \widetilde{F}||_\infty < \varepsilon$. Let

$$G(x) = \left(1 - \sum_{j=1}^l y_j\right)\widetilde{G}(x) + \sum_{j=1}^l y_j \mathbb{1}\{x \geq x_j\}.$$

Then $G \in \cup_{k=1}^{\infty}\mathcal{P}_{m_0}^k$ and $||G-F||_\infty < \varepsilon$. Consequently, we have shown that $\cup_{k=1}^{\infty}\mathcal{P}_{m_0}^k$ is dense in $\mathcal{P}_{m_0}$.

Now, let $\varepsilon > 0$ and $F_\varepsilon \in \mathcal{P}_{m_0}$ be such that $\overline{\theta}_0 \leq g(F_\varepsilon) + \varepsilon$. By continuity of g, there exists δ such that $||F - F_\varepsilon|| < \delta$ implies that $|g(F) - g(F_\varepsilon)| < \varepsilon$. By what precedes, there also exists k_0 and $F_{k_0,\varepsilon} \in \mathcal{P}^{k_0}_{m_0}$ such that $||F_{k_0,\varepsilon} - F_\varepsilon|| < \delta$. As a result, $\overline{\theta}_0 \leq g(F_{k_0,\varepsilon}) + 2\varepsilon$. Hence, for all $k \geq k_0$, $\overline{\theta}^k \leq \overline{\theta}_0 \leq \overline{\theta}^k + 2\varepsilon$ since $\mathcal{P}^{k_0}_{m_0} \subset \mathcal{P}^k_{m_0}$. This proves Part 1 of the theorem.

2. $\overline{\theta}_0 = \overline{\theta}^{K+1}$ for g convex.

We prove Part 2 in three steps.

a. For all k, the supremum of g on $\mathcal{P}^k_{m_0}$ is reached on an extremal element of this set.

First, because $\mathcal{P}^k_{m_0}$ is closed and bounded and belongs to a finite dimensional space, it is compact. Then by continuity of $g(.)$,

$$\overline{\theta}^k = \max_{F \in \mathcal{P}^k_{m_0}} g(F).$$

Let $F_0 \in \arg\max_{F \in \mathcal{P}^k_{m_0}} g(F)$. If F_0 is an extremal point, step a is proved. Otherwise, by Minkowksi's and Caratheodory's theorems (see, e.g., Hiriart-Urruty & Lemaréchal, 2001, Theorems 2.3.4 and 1.3.6 respectively), there exists an extremal distribution F_1, $\lambda \in (0, 1]$ and $F_2 \in \mathcal{P}^k_{m_0}$ such that

$$F_0 = \lambda F_1 + (1 - \lambda)F_2.$$

Because g is convex,

$$g(F_0) \leq \lambda g(F_1) + (1 - \lambda)g(F_2) \leq \lambda g(F_1) + (1 - \lambda)g(F_0).$$

Thus, $g(F_1) \geq g(F_0)$ and the result follows by definition of F_0.

b. Extremal points of $\mathcal{P}^k_{m_0}$ $(k > K + 1)$ belong to $\mathcal{P}^{K+1}_{m_0}$.

We shall prove that distributions with n' points of support, with $k \geq n' > K + 1$, are not extremal points of $\mathcal{P}^k_{m_0}$. Let F_0 be such a distribution, $(x_1, ..., x_{n'})$ be its support points and $(y_1, ..., y_{n'})$ be the associated probabilities. The vector $m_0 \in \mathbb{R}^K$ belongs to the convex hull of the set $A = \{(x_1, ..., x_1^K)', ..., (x_{n'}, ..., x_{n'}^K)'\}$. Thus, by Carathéodory's theorem, there exists $K + 1$ points in A (say, $(x_1, ..., x_{K+1})$) and $(q_1, ..., q_{K+1}) \in [0, 1]^{K+1}$ such that $m_k = \sum_{j=1}^{K+1} q_j x_j^k$ for all $1 \leq k \leq K$ and $\sum_{j=1}^{K+1} q_j = 1$. In other words, we have defined a distribution function F_1, with support $(x_1, ..., x_{K+1})$ and associated probabilities $(q_1, ..., q_{K+1})$ such that $F_1 \in \mathcal{P}^{K+1}_{m_0}$. Now let $0 < \lambda < \min_{i=1...n'}(\min(y_i/q_i, (1 - y_i)/(1 - q_i), 1/2))$ (where we let $q_i = 0$ for $i > K + 1$ and $t/0 = +\infty$ for any $t > 0$) and define, for $i = 1...n'$,

$$r_i = \frac{y_i - \lambda q_i}{1 - \lambda}.$$

Then it is easy to see that $r_i \in [0,1]$ for all i. Let F_2 denote the distribution function with support points $(x_1, ..., x_{n'})$ and associated probabilities $(r_1, ..., r_{n'})$. Then $F_2 \in \mathcal{P}^k_{m_0}$. Besides, by construction,

$$F_0 = \lambda F_1 + (1-\lambda)F_2.$$

Thus F_0 is not an extremal element of $\mathcal{P}^k_{m_0}$. As a result, extremal points of $\mathcal{P}^k_{m_0}$ are distributions with at most $K+1$ points of support .

c. $\overline{\theta}_0 = \overline{\theta}^{K+1}$.

To finish the proof, let $\varepsilon > 0$. By Part 1, there exists $k_0 > K+1$ such that $\overline{\theta}_0 \leq \overline{\theta}^{k_0}_0 + \varepsilon$. By steps a and b, $\overline{\theta}^{k_0}_0 \leq \overline{\theta}^{K+1}_0$. Thus, $\overline{\theta}_0 \leq \overline{\theta}^{K+1}_0 + \varepsilon$. This shows that $\overline{\theta}_0 \leq \overline{\theta}^{K+1}_0$ because ε was arbitrary. The result follows, since the other inequality trivially holds $\square$

Proposition 2.3

For any increasing and concave function u, by Jensen's inequality,

$$\begin{aligned} E\left[u\left(\frac{X}{K}\right)\right] &= E\left[E\left[u\left(\frac{X}{K}\right)\middle|p\right]\right] \\ &\leq E\left[u\left(E\left[\frac{X}{K}\middle|p\right]\right)\right] \\ &\leq E[u(p)]. \end{aligned}$$

Hence, F_p dominates stochastically $F_{\frac{X}{K}}$ at the second order, and by monotonicity, $g(F_p) \leq \widetilde{\theta}$. Moreover, this is true for any distribution $F_p \in \mathcal{P}_{m_0}$ since such distributions rationalize the one of $\frac{X}{K}$. Choosing a sequence $(F_{n,p})_{n\in\mathbb{N}}$ in $\mathcal{P}_{m_0}$ such that $\lim_{n\to\infty} g(F_{n,p}) = \overline{\theta}_0$, we thus get $\overline{\theta}_0 \leq \widetilde{\theta}$. When the support of p is not reduced to $\{0,1\}$, $\frac{X}{K}$ is not a deterministic function of p with probability equal to one. Hence, for any strictly concave function u, the event $E\left[u\left(\frac{X}{K}\right)|p\right] < u\left(E\left[\frac{X}{K}|p\right]\right)$ holds with a positive probability. As a result, $E\left[u\left(\frac{X}{K}\right)\right] < E[u(p)]$, and the result follows by strict monotonicity of g $\square$

Proposition 3.1

By definition, $\widehat{m} \in \mathcal{M}$. Thus, the first part follows by, for instance Theorem 4.1 and 4.3 of by Curto & Fialkow (1991). As for the second part, define

$$\mathcal{B}_K = \{(\Pr(X=0|p=u), ..., \Pr(X=K|p=u))', u \in [0,1]\},$$

and let $\text{co}(\mathcal{B}_K)$ denote the convex hull of $\mathcal{B}_K$. Then observe that $\widetilde{m} \in \mathcal{M}$ if and only if $(\widehat{P}_0, ..., \widehat{P}_K)' \in \text{co}(\mathcal{B}_K)$. Then Wood (1999) shows that when $(\widehat{P}_0, ..., \widehat{P}_K)' \notin \text{co}(\mathcal{B}_K)$, the projection of $(\widehat{P}_0, ..., \widehat{P}_K)'$ onto this set with respect to any Euclidian norm corresponds to a unique distribution for p with at most $L+1$ points of support (see Wood, 1999, Theorem 5.1). Thus, to prove the result, it suffices to show that $Q\widehat{m}_W$ corresponds to the projection of $(\widehat{P}_0, ..., \widehat{P}_K)'$ onto $\text{co}(\mathcal{B}_K)$ with respect to an appropriate norm.

We first define this norm. Let e denote the vector of ones of size K and let $0 < c < 2/(e'QW^{-1}Q'e)$. Define

$$\Lambda = \left(\begin{array}{c|ccc} 2c & c & \dots & c \\ \hline c & & & \\ \vdots & & Q'^{-1}WQ^{-1} & \\ c & & & \end{array}\right).$$

Then Λ is positive definite because both $Q'^{-1}WQ^{-1}$ and its Schur complement $2c - c^2e'QW^{-1}Q'e$ are positive definite. Moreover, letting $||.||_\Lambda$ denote the norm in $\mathbb{R}^{K+1}$ induced by Λ, we have, for any $x = (x_0, ..., x_K), y = (y_0, ..., y_K)$,

$$||x - y||_\Lambda = ||x_{-0} - y_{-0}||_{Q'^{-1}WQ^{-1}}, \tag{7.1}$$

where $x_{-0} = (x_1, ..., x_K)$ and similarly for y_{-0}.

Now, let us consider the projection $(\widehat{P}_{\Lambda,0}, ..., \widehat{P}_{\Lambda,K})'$ of $(\widehat{P}_0, ..., \widehat{P}_K)$ onto $\text{co}(\mathcal{B}_K)$ with respect to the norm induced by Λ. Using (7.1) and the definition of $\widehat{m}_W$, we have

$$\begin{aligned}
(\widehat{P}_{\Lambda,1}, ..., \widehat{P}_{\Lambda,K})' &= \arg\min_{x\in\mathbb{R}^{K+1}} \left\|(\widehat{P}_0, ..., \widehat{P}_K)' - x\right\|_\Lambda \\
&= \arg\min_{x\in\mathbb{R}^{K}} \left\|\widehat{P} - x\right\|_{Q'^{-1}WQ^{-1}} \\
&= \arg\min_{x\in\mathbb{R}^{K}} \|Q\widetilde{m} - x\|_{Q'^{-1}WQ^{-1}} \\
&= Q\arg\min_{x\in\mathbb{R}^{K}} \|\widetilde{m} - x\|_W \\
&= Q\widehat{m}_W \ \square
\end{aligned}$$

Theorem 3.2

Because $m_0 \in \overset{\circ}{\mathcal{M}}$, $\Pr(\widetilde{m} \in \overset{\circ}{\mathcal{M}}) \to 1$ and $\widehat{m}$ has the same asymptotic distribution as $\widetilde{m}$:

$$\sqrt{n}\,(\widehat{m}_W - m_0) \xrightarrow{d} \mathcal{N}(0, \Sigma). \tag{7.2}$$

Assumption 3.1 ensures that we can apply the smooth maximum theorem (see, e.g., Carter, 2001, Theorem 6.1). As a result, $\overline{\theta}^k$ and $\underline{\theta}^k$ are C^1 on a neighborhood of m_0. Moreover, by the smooth envelope theorem (see Carter, 2001, Corollary 6.1.1),

$$\frac{\partial\overline{\theta}^k}{\partial m}(m_0) = \frac{\partial q^k}{\partial m}(\overline{x}_0, m_0) - \overline{y}_0'A(\overline{x}_0)'\overline{\lambda}_0.$$

where $\overline{\lambda}_0$ is the vector of Lagrange multiplier corresponding to the first order condition of (3.3). The same holds for the lower bound. Then the asymptotic normality and the expression of the asymptotic variance follow by (7.2), the delta method and the Cramér-Wold device. Finally, because $\overline{\theta}^k$ and $\underline{\theta}^k$ are C^1, $\frac{\partial \overline{\theta}^k}{\partial m}(\widehat{m})$ and $\frac{\partial \underline{\theta}^k}{\partial m}(\widehat{m})$ are consistent estimators of $\frac{\partial \overline{\theta}^k}{\partial m}(m_0)$ and $\frac{\partial \underline{\theta}^k}{\partial m}(m_0)$. By the weak law of large numbers, $\widehat{\Sigma} \xrightarrow{P} \Sigma$. As a result, $\widehat{J}'\widehat{\Sigma}\widehat{J} \xrightarrow{P} J'\Sigma J$ □

Corollary 3.3

By Part 3 of Theorem 2.2, $\overline{\theta}^{K+1} = \overline{\theta}_0$ and $\underline{\theta}^{K+1} = \underline{\theta}_0$. By consistency of $\widehat{\overline{\theta}}^{K+1}$ and $\widehat{\underline{\theta}}^{K+1}$, $Pr(\theta \in \text{CI}^1_{1-\alpha}) \to 1$ for all $\theta \in (\underline{\theta}_0, \overline{\theta}_0)$. Besides, by Theorem 3.2,

$$\lim_{n\to\infty} Pr(\underline{\theta}_0 \in \text{CI}^1_{1-\alpha}) = \lim_{n\to\infty} Pr(\overline{\theta}_0 \in \text{CI}^1_{1-\alpha}) = 1 - \alpha.$$

The result follows □

Corollary 3.4

Remark that

$$\begin{aligned} m_0 \in I_{1-\alpha} \quad &\Rightarrow \quad \underline{\theta}^{K+1}(m_0) \in \underline{\theta}^{K+1}(I_{1-\alpha}),\ \overline{\theta}^{K+1}(m_0) \in \overline{\theta}^{K+1}(I_{1-\alpha}) \\ &\Rightarrow \quad [\underline{\theta}_0, \overline{\theta}_0] \subset \left[\inf_{m\in I_{1-\alpha}} \underline{\theta}^{K+1}(m),\ \sup_{m\in I_{1-\alpha}} \overline{\theta}^{K+1}(m)\right], \end{aligned}$$

where the second implication follows by Part 3 of Theorem 2.2. The result follows by definition of $I_{1-\alpha}$ □

References

Allen, R., Burgess, S. & Windmeijer, F. (2009), More Reliable Inference for Segregation Indices. University of Bristol Working Paper No 09/216.

Aslund, O. & Skans, O. N. (2010), 'Will I See You at Work? Ethnic Workplace Segregation in Sweden, 1985-2002', *Industrial and Labor Relations Review* **63**(3), 471–493.

Bayard, K., Hellerstein, J., Neumark, D. & Troske, K. (1999), Why Are Racial and Ethnic Wage Gaps Larger for Men than for Women? Exploring the Role of Segregation Using the New Worker-Establishment Characteristics Database, *in* Haltiwanger, Lane, Spletzer, Theeuwes & Troske, eds, 'The Creation and Analysis of Employer-Employee Matched Data', Elsevier Science B.V. (Amsterdam), pp. 175–203.

Carrington, W. J. & Troske, K. R. (1995), 'Gender Segregation in Small Firms', *Journal of Human Resources* **30**(3), 503–533.

Carrington, W. J. & Troske, K. R. (1997), 'On Measuring Segregation in Samples with Small Units', *Journal of Business & Economic Statistics* **15**(4), 402–09.

Carrington, W. J. & Troske, K. R. (1998*a*), 'Interfirm Segregation and the Black/White Wage Gap', *Journal of Labor Economics* **16**(2), 231–60.

Carrington, W. J. & Troske, K. R. (1998*b*), 'Sex Segregation in U.S. Manufacturing', *Industrial and Labor Relations Review* **51**(3), 445–464.

Carter, M. (2001), *Foundations of Mathematical Economics*, MIT Press.

Chernozhukov, V., Fernandez-Val, I., Hahn, J. & Newey, W. (2009), Identification and Estimation of Marginal Effects in Nonlinear Panel Data Models. Working paper.

Cortese, C. F., Falk, F. & Cohen, J. (1978), 'Understanding the Standardized Index of Dissimilarity: Reply to Massey', *American Sociological Review* **43**(4), 590–592.

Cortese, C., Falk, F. & Cohen, J. K. (1976), 'Further Considerations on the Methodological Analysis of Segregation Indices', *American Sociological Review* **41**(4), 630–637.

Curto, R. E. & Fialkow, L. A. (1991), 'Recursiveness, Positivity, and Truncated Moment Problems', *Houston Journal of Mathematics* **17**, 603–635.

Dustmann, C., Glitz, A. & Schönberg, U. (2010), Referral-based Job Search Networks. mimeo UCL.

Fields, J. & Wolff, E. N. (1991), 'The Decline of Sex Segregation and the Wage Gap, 1970-80', *Journal of Human Resources* **26**(4), 608–622.

Giuliano, L., Levine, D. I. & Leonard, J. (2009), 'Manager Race and the Race of New Hires', *Journal of Labor Economics* **27**(4), 589–631.

Groshen, E. L. (1991), 'The Structure of the Female/Male Wage Differential: Is It Who You Are, What You Do, or Where You Work?', *Journal of Human Resources* **26**(3), 457–472.

Hellerstein, J. K. & Neumark, D. (2008), 'Workplace Segregation in the United States: Race, Ethnicity, and Skill', *The Review of Economics and Statistics* **90**(3), 459–477.

Hiriart-Urruty, J.-B. & Lemaréchal, C. (2001), *Fundamentals of Convex Analysis*, Springer.

Imbens, G. W. & Manski, C. (2004), 'Confidence Intervals for Partially Identified Parameters', *Econometrica* **72**, 1845–1857.

Insee, ed. (2005), *Les immigrés en France*, Collection Références, 2005 edn, Insee.

Iranzo, S., Schivardi, F. & Tosetti, E. (2008), 'Skill Dispersion and Firm Productivity: An Analysis with Employer-Employee Matched Data', *Journal of Labor Economics* **26**(2), 247–285.

Kramarz, F., Lollivier, S. & Pelé, L.-P. (1996), 'Wage Inequalities and Firm-Specific Compensation Policies in France', *Annales d'Economie et de Statistique* **41-42**, 369–386.

Krein, M. G. & Nudel'man, A. A. (1977), *The Markov Moment Problem and Extremal Problems*, Translations of Mathematical monographs.

Kremer, M. & Maskin, E. (1996), Wage Inequality and Segregation by Skill, NBER Working Papers 5718, National Bureau of Economic Research, Inc.

Lord, F. M. (1969), 'Estimating True-Score Distributions in Psychological Testing (an Empirical Bayes Estimation Problem)', *Psychometrika* **34**, 259–299.

Pistaferri, L. (1999), 'Informal Networks in the Italian Labor Market', *Giornale degli Economisti* **58**(3-4), 355–375.

Rathelot, R. (2011), Measuring Segregation When Units Are Small: a Parametric Approach. Crest Working Paper.

Söderström, M. & Uusitalo, R. (2010), 'School Choice and Segregation: Evidence from an Admission Reform', *Scandinavian Journal of Economics* **112**(1), 55–76.

Winship, C. (1977), ‘A Revaluation of Indexes of Residential Segregation’, *Social Forces* **55**(4), 1058–1066.

Wood, G. R. (1999), ‘Binomial mixtures: Geometric Estimation of the Mixing Distribution’, *Annals of Statistics* **27**, 1706–1721.

Wooldridge, J. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.