ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

D'Haultfoeuille, Xavier; Rathelot, Roland

Working Paper

Measuring Segregation on Small Units: A Partial Identification Analysis

CReAM Discussion Paper Series, No. 11/16

Provided in Cooperation with:
Rockwool Foundation Berlin (RF Berlin)

Suggested Citation: D'Haultfoeuille, Xavier; Rathelot, Roland (2016) : Measuring Segregation on Small Units: A Partial Identification Analysis, CReAM Discussion Paper Series, No. 11/16, Centre for Research & Analysis of Migration (CReAM), Department of Economics, University College London, London

This Version is available at:
https://hdl.handle.net/10419/295538

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

CReAM
Centre for Research & Analysis of Migration

Discussion Paper Series

CPD 11/16

Measuring Segregation on Small Units: A Partial Identification Analysis

Xavier D'Haultfoeuille and Roland Rathelot

Centre for Research and Analysis of Migration
Department of Economics, University College London
Drayton House, 30 Gordon Street, London WC1H 0AX

www.cream-migration.org

Measuring Segregation on Small Units: A Partial Identification Analysis*

Xavier D'Haultfœuille† and Roland Rathelot‡

April 2016

Abstract

We consider the issue of measuring segregation in a population of small units, considering establishments in our application. Each establishment may have a different probability to hire an individual from the minority group. We define segregation indices as inequality indices on these unobserved, random probabilities. Because these probabilities are measured with error by proportions, standard estimators are inconsistent. We model this problem as a nonparametric binomial mixture. Under this testable assumption and conditions satisfied by standard segregation indices, such indices are partially identified and sharp bounds can be easily obtained by an optimization over a low dimensional space. We also develop bootstrap confidence intervals and a test of the binomial mixture model. Finally, we apply our method to measure the segregation of foreigners in small French firms.

JEL Classification: C13, C14 and J71.

Keywords: segregation, small units, partial identification.

*We would like to thank the editor, two anonymous referees, Romain Aeberhardt, Mathias André, Christian Bontemps, Clément de Chaisemartin, Laurent Davezies, Guy Laroque, Adam Rosen and the participants of the ESEM in Oslo, 21st EC^2 conference in Toulouse, as well as seminars at Crest, INSEE and Lausanne. The access to confidential data was partly funded by the grant ANR-10-EQPX-17.

†CREST, xavier.dhaultfoeuille@ensae.fr.

‡University of Warwick, r.rathelot@warwick.ac.uk.

1 Introduction

Suppose that we seek to measure to what extent a minority group, such as foreigners, is concentrated in some firms only, because for instance some firms are reluctant to hire them.[1] Measuring the magnitude of segregation is a crucial step to understand the underlying phenomena and design adequate policies.[2] A natural way to do this would be to compute the proportion of minority workers X_i/K_i, with X_i the number of workers from the minority group and K_i the size of firm i, and then compute an inequality index on the sample $(X_i/K_i)_{i=1...n}$. However, even if firms all hire each of their worker from the minority group with the same probability, there will be some variation on X_i/K_i across firms just because K_i is finite. Hence, this approach will overestimate the actual level of segregation, an issue known as the *small-unit bias*.

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. The idea was developed by Carrington and Troske (1997), who propose an adjustment that can be applied to other indices. Allen et al. (2015) proposed a correction based on bootstrap. These papers have received substantial attention, as the literature on the measurement of segregation, whether at the residential (see e.g. Cutler and Glaeser, 1997; Cutler et al., 1999, 2008; Echenique and Fryer, 2007; Bayer and McMillan, 2012), school (see e.g. Card and Rothstein, 2007; Fredriksson et al., 2013) or workplace level (see e.g. Carrington and Troske, 1998; Hellerstein and Neumark, 2008) is large. Having only a small number of observations per unit is particularly frequent for workplace and school segregation as a large share of firms have less than ten employees and classrooms are between 20 and 40 pupils.[3] Residential segregation may also be affected when only surveys (and not censuses) are available. In their paper about ideological segregation of the Internet, Gentzkow and Shapiro (2011) have also to deal with small-unit bias.

[1] Hereafter, we illustrate our ideas with the example of firms, in line with our application, but of course they also apply to different units, in particular geographic areas and classrooms.

[2] All along the paper, we follow the literature and use the term "segregation" in a positive term, to describe the relative concentration of groups across units. The indices we propose do not allow, by themselves, to conclude about which segregating mechanism is at work.

[3] See also Söderström and Uusitalo (2010); Brunello and Rocco (2013); Leckie and Goldstein (2014) for recent papers about school segregation with attempts to correct for small-unit bias.

In this paper, we propose a different approach: we consider that segregation should be measured directly through an inequality index on the distribution F_p of p_i, the probability that firm i has to hire someone from the minority group.[4] In line with the literature (see in particular Winship, 1977; Carrington and Troske, 1997; Rathelot, 2012), we impose an independence condition between hiring or, more generally, on the allocation process. In other words, conditional on K_i and p_i, X_i is supposed to follow a binomial distribution $B(K_i, p_i)$. This binomial assumption, which we show is testable, allows us to identify the first moments of F_p. 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 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 linearity condition satisfied by, among others, the Atkinson, Duncan and Theil indices, the bounds on segregation indices for units of size K can be obtained by optimizing over discrete distributions with only $K + 1$ points of support at most. We also show, using the theory of Chebyshev systems (see, e.g. Krein and Nudel'man, 1977) that under another assumption satisfied for instance by the Theil index, bounds can be obtained without optimization, by simply finding roots of appropriate polynomials. We also show how bounds on subpopulations can be combined to handle random unit sizes or to control for covariates at the unit (e.g., firms' sectors) or at the position (e.g., skilled versus unskilled positions in the firm) level.

Our results are also related to a few papers on partial identification. Stoye (2010) also considers the partial identification of spread parameters such as the Gini, Theil or Atkinson indices. Our paper complements his, by considering a different type of measurement problem on the variable of interest p. While Stoye (2010) considers missing or interval valued data, we consider a setting where only restrictions on the first moments of p are available. Our identification result is also linked to a result of Chernozhukov et al. (2013) in the context of nonlinear panel data models. In such models, bounds on marginal effects can be obtained by maximizing some functionals over the distribution of the fixed effect. Similarly to

[4] In Section 2, we elaborate on when one should focus on F_p rather than on the distribution of the realized share $(X_i/K_i)_{i=1\dots n}$.

us, they show that one can actually restrict to discrete distributions with a low number of support points.

We also develop estimation and inference on the segregation index, using a two-step procedure. In the first step, we consider a maximum likelihood estimator for the distribution of X_i conditional on K_i. Once the unobserved p_i is integrated out, X_i does not follow a binomial distribution in general, but a multinomial one, with some inequality constraints on the corresponding probabilities stemming from the underlying binomial model. The estimator takes a simple closed form when the constraints are slack. When they are not, we show that the estimator can be obtained through an optimization under 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. When the constraint on the vector of moments is binding, the lower and upper bounds coincide and no optimization is needed in this case. We show that the estimated bounds are consistent under minimal conditions and derive their asymptotic distribution under additional restrictions. This distribution is normal when the true vector of moments lies in the interior of the moment space, but is not when this vector is at the boundary of the moment space. We propose a bootstrap confidence interval that works in both cases. Finally, we develop a bootstrap likelihood ratio test of the binomial mixture model.

Monte Carlo simulations indicate that our method works well for finite samples, and is not computationally too demanding. They also show that even for modest unit sizes ($K = 9$, 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. For typical unit and sample sizes, the length of the confidence intervals mostly stems from sampling variation, not from partial identification.

Finally, we apply our framework to measure the segregation of immigrants in small French firms. Our method proves to work well in this context. First, we do not reject the binomial mixture model for any plant sizes. Second, for plant sizes larger than 3, the identification region is already informative. Third, contrary to what is suggested by the naive, Carrington and Troske (1997) or Allen et al. (2015) estimators, we cannot reject at standard levels that there is no relationship between plant size and the level of segregation, at least for very small firms.

Finally, we show that the level of segregation we obtain is not explained by the fact that foreigners hold more unskilled positions. This result is consistent with those of Hellerstein and Neumark (2008) and Åslund and Skans (2010) for American and Swedish firms respectively.

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 presents the estimation procedure for the bounds, as well as the inference results. The behavior of the bounds and their estimators is studied through simulations in Section four. The application to workplace segregation is developed in the fifth section. Section six concludes. Appendix A gathers all the proofs. In the online appendix, we extend our framework to handle random unit sizes and control for covariates, develop the bootstrap likelihood ratio test of the binomial mixture model and provide additional discussion on inference.

2 Identification

2.1 The setting and the object of interest

The population is assumed to be split into two groups, a group of interest, called the minority group hereafter, and the rest of the population. Individuals are distributed across units, which may represent geographic areas, classrooms, or, as in our application, firms. We assume that there exists a random variable p_i taking values in $[0, 1]$ that represents the probability for any individual belonging to unit i to be a member of the population of interest. The probabilities p_i are i.i.d. across units, with cumulative distribution function (cdf) F_p. Because we have in mind units of small to moderate size, our asymptotic analysis here is in the number of units.[5]

The object of interest of this paper is a segregation index of the minority group, θ_0, which is a real functional of F_p. We denote hereafter this index by $g(F_p, m_{01})$, with $m_{01} = E(p)$. This notation may seem redundant, because m_{01} depends on F_p, but the reason will become clearer below. Popular indices include the Duncan D, the Theil T, the Atkinson A_b parameterized by $b \in (0, 1)$, and the coworker

[5]This contrasts with the framework of Allen et al. (2015) where the size of the units tend to infinity while the number of units is fixed.

index CW. They satisfy respectively

$$\begin{aligned} D &= \tfrac{1}{2}E\left[\left|\tfrac{p}{E(p)} - \tfrac{1-p}{E(1-p)}\right|\right] = \tfrac{\int |u-m_{01}|dF_p(u)}{2m_{01}(1-m_{01})}, \\ T &= 1 - \tfrac{E(p\ln(p))}{E(p)\ln(E(p))} = 1 - \tfrac{\int u\ln(u)dF_p(u)}{m_{01}\ln(m_{01})}, \\ A_b &= 1 - m_{01}^{-\frac{b}{1-b}}(1-m_{01})^{-1}\left(\int (1-u)^{1-b}u^b dF_p(u)\right)^{\frac{1}{1-b}}, \\ CW &= \int (u-m_{01})^2 dF_p(u)/(m_{01}-m_{01}^2). \end{aligned} \tag{2.1}$$

These definitions correspond to the limit in probability of the standard formulas given for instance by Massey and Denton (1988). Though our main results will not include it, let us also mention the Gini index, $G = (1-m_{01}-\int F_p^2(u)du)/m_{01}$.

The probability p are not directly observed. Instead, we observe the size of the unit, K, and the number X of minority individuals in that unit.[6] Why is it interesting to learn about F_p rather than about the distribution of the realized share X/K or of the number of minority individuals across units? While the realized shares may be more interesting for studying the consequences of segregation, as we discuss in the conclusion, we see at least two reasons for this choice.

First, let us consider the case where the available data are not exhaustive: suppose that only a subset of individuals in a school, a neighborhood or a firm are sampled. In this case, $F_{X/K}$, the distribution of the realized minority shares, is not observed. If one is interested in this distribution, we show that our analysis can be applied, up to a few changes detailed in section 2.2. When the size of the underlying unit is very large (but the sample size remains small), our results can be applied as such.

Second, since the beginning of the segregation literature, computing segregation indices is often used to understand the features of the underlying allocation process of individuals across units. If the question is to investigate whether the allocation process, as a whole, is influenced by the ethnicity variable, the interesting distribution is less the realized one (which, by construction, incorporates the noise coming from the sampling process) than the one of the underlying probabilities. In one of the earliest paper of this literature, Jahn et al. (1947) characterizes in this way the absence of segregation: "[...] if there is no segregation then members of a minority racial group [...] will be distributed randomly throughout the various census tracts of a city".

[6] To ease the exposition, we omit subscripts i in this section.

To fix ideas, consider the following model. In a first step, a job seeker has to choose to which firm i to apply to. In a second step, firm i decides which job seekers they hire. To make their decision, job seekers will consider the nature of the job, the wage offered, as well as the distance between the job and his residence. Firms will try to assess candidates' productivity based on observable information; they can also be prejudiced. The probability p_i of firm i will then depend of both the probability to apply of minority and majority job seekers, and, conditional on the application, on the probability for the firm to hire the majority or minority applicant. If at both stages ethnicity is not relevant, that is if minority and majority apply to similar jobs and employers make similar hiring decisions when they face minority and majority candidates, we expect p_i to reflect the proportion of job seekers in the population, so that $p_i = m_{01}$ for all i and $D = T = A_b = CW = 0$. However, given the small number of workers in firms, the actual proportion of minority workers may only be a poor approximation of p_i, and the segregation indices based on the distribution of X/K instead of p would be all positive, even though ethnicity would play no role in the underlying process.[7]

From this example, it is obvious that computing segregation based on the distribution of p does not exhaust the set of interesting questions relating to the underlying process. One of the most important caveat is that the segregation measured with respect to ethnicity may well be driven by any characteristics correlated with ethnicity. For instance, if minority and majority workers differ in their skill level, ethnic segregation may just be due to the fact that firms vary in the level of skill that they require.This issue is not specific to the small-unit case and also applies when units are large. In the online appendix, we propose a solution by including covariates at the unit level (e.g., firms' sector) or at the position level (e.g., skilled vs unskilled positions). Another limit is that applications and hiring decisions may depend on the current, realized share of the minority in the firm. If so, realized shares may be of more interest than the unobserved probabilities. But such a case would also violate the binomial mixture model we consider below. To assess whether this issue is important in a given context, it is therefore important to run the test of the binomial mixture model that we develop

[7] A similar reasoning would apply to education. School and classroom segregation results from both the geographic distribution of ethnic groups, individual choices and from principals decisions in accepting pupils and gathering them into classrooms.

in Subsection C.1 of the online appendix.

2.2 The main identification result

We suppose here that the size of units K is constant; the case of a random size is considered in Section B.1 of the online appendix. We posit that individuals are selected into units independently from each other 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 and Troske (1997) or Rathelot (2012). It is also asymptotically equivalent to the allocation mechanism considered by Allen et al. (2015) when the number of individuals and the number of units tend to infinity at the same rate.[8] Finally, as we shall see below, this assumption is testable.

Because the distribution of X is defined by K probabilities, namely $P_0 = (P_{01}, ..., P_{0K})'$, with $P_{0j} = \Pr(X = j)$, we expect it to convey information on K parameters of F_p. Letting $m_{0i} = E(p^i)$, we have, after some algebra:

$$P_{0j} = E\left[\Pr(X = j|p)\right] = \sum_{i=1}^{K} \binom{K}{i} \binom{i}{j} (-1)^{i-j} m_{0i},$$

Hence, letting $m_0 = (m_{01}, ..., m_{0K})'$ and Q be the $K \times K$ matrix of typical element $\binom{K}{j}\binom{j}{i}(-1)^{j-i}$, we get

$$P_0 = Q m_0. \tag{2.2}$$

Moreover, Q is invertible as an upper triangular matrix with non-zero diagonal elements. Thus, there is a one-to-one mapping between P_0 and m_0. This has two implications. First, m_0 is identified from the distribution of X. As a result, any

[8] Allen et al. (2015) suppose that individuals from group $e \in \{0, 1\}$ (with $e = 1$ for the minority group, say) are allocated independently and with probability π_i^e to unit i. However, if the number of individuals n tends to infinity together with the number of units, such that $\pi_i^e n \to \rho_i^e$, then the number of individuals of group e in unit i, X_i^e, follows a Poisson distribution with parameter ρ_i^e. Because X_i^1 and X_i^0 are independent, we finally get $X_i^1 | X_i^1 + X_i^0 = K \sim B(K, \rho_i^1/(\rho_i^0 + \rho_i^1))$, as here.

parameter θ_0 depending only on m_0 is point identified. This is for instance the case of the coworker index CW. Because $CW = (m_{02} - m_{01}^2)/(m_{01} - m_{01}^2)$, the coworker index is point identified as soon as $K \geq 2$.

The second implication of (2.2) is that two different distributions of p with the same first K moments lead to the same distribution of X and are thus observationally equivalent. In other words, we do not learn anything on p beyond its first K moments. As a result, θ_0 is not identified in general, and its sharp lower and upper bounds $\underline{\theta}_0$ and $\overline{\theta}_0$ satisfy

$$\underline{\theta}_0 = \inf_{F \in \mathcal{D}_{m_0}} g(F, m_{01}), \quad \overline{\theta}_0 = \sup_{F \in \mathcal{D}_{m_0}} g(F, m_{01}), \tag{2.3}$$

where $\mathcal{D}_{m_0}$ is the subset of $\mathcal{D}$, the set of cumulative distribution functions on $[0, 1]$, for which the vector of first K moments equals m_0.

Equation (2.3) provides the sharp bounds on θ_0 but is not useful in practice because it amounts to optimizing over an infinite dimensional set. We now show that under restrictions satisfied by most segregation indices, the problem can be much simplified. We use for that purpose related results on the so-called Chebyshev-Markov moment problem (see, e.g., Krein and Nudel'man, 1977, for historical notes on this problem).

As a vector of raw moments, m_0 cannot lie anywhere in $[0, 1]^K$. It should satisfy some restrictions; for instance, the variance has to be positive, implying $m_{02} \geq m_{01}^2$. Formally,

$$m_0 \in \mathcal{M} = \left\{ \left(\int x dF, ..., \int x^K dF \right)', \ F \in \mathcal{D} \right\}.$$

We provide a complete characterization of the moment space $\mathcal{M}$ below, but first consider the case where m_0 belongs to its boundary $\partial\mathcal{M}$. When $m_0 \in \partial\mathcal{M}$, there is actually a unique distribution F^* corresponding to m_0. Moreover, F^* is discrete with at most $L + 1$ support points, where L is the integer part of $(K + 1)/2$ (for a proof of both points, see, e.g., Theorem IV.4.1 in Krein and Nudel'man, 1977). Then no optimization is required to solve (2.3), and $\underline{\theta}_0 = \overline{\theta}_0$.

Now, when $m_0 \in \overset{\circ}{\mathcal{M}}$, the interior of $\mathcal{M}$, we can also simplify the computation of the bounds, at the price of imposing the following assumption.

Assumption 2.1 *$g(F, m_{01}) = \nu\left(\int h(x, m_{01}) dF(x), m_{01}\right)$, where h and ν are continuous and $\nu(., m_{01})$ is monotonic.*

An important feature of the assumption is that $F \mapsto \int h(x, m_{01})dF(x)$ is linear. Assumption 2.1 does not hold for the Gini index but is satisfied by the Duncan, the Theil and the Atkinson indices.[9]

Assumption 2.1 also holds when one cares about the realized shares, but only a sample of each unit is observed. Specifically, let $\widetilde{X}$ denote the total number of minority people in a random unit of size L, and suppose that the parameter of interest satisfies $\theta_0 = \nu\left(\int h(x, m_{01})dF_{\widetilde{X}/L}(x), m_{01}\right)$. Assume that in each unit, only $K < L$ individuals are sampled, among whom X belong to the minority. Then, using the fact that $\widetilde{X} - X|p, X \sim B(p, L-K)$, we obtain, after some algebra,

$$\theta_0 = \nu\left(\int \widetilde{h}(p, m_{01})dF_p(p), m_{01}\right),$$

with

$$\widetilde{h}(p, m_{01}) = \sum_{j=0}^{K}\sum_{k=0}^{L-K}\binom{K}{j}\binom{L-K}{k}h((j+k)/K, m_{01})p^{j+k}(1-p)^{L-(j+k)}.$$

Hence, Assumption 2.1 also holds in this context. Note that when $L \to \infty$, the law of large numbers and continuity of h imply that $\widetilde{h}(p, m_{01}) \to h(p, m_{01})$.

Under this condition, by a theorem of Caratheodory, the bounds on $\int h(x, m_{01})dF(x)$, and thus on θ_0, are attained on distributions with no more than $K+1$ support points (see for instance Theorem I.3.6 of Krein and Nudel'man, 1977).[10] This makes the optimization computationally possible. Specifically, let $\mathcal{D}^\ell$ denote the subset of $\mathcal{D}$ with at most ℓ points of support and $\mathcal{D}^\ell_{m_0} = \mathcal{D}^\ell \cap \mathcal{D}_{m_0}$. Then define

$$\underline{\theta}_{0,\ell} = \inf_{F \in \mathcal{D}^\ell_{m_0}} g(F, m_{01}), \quad \overline{\theta}_{0,\ell} = \sup_{F \in \mathcal{D}^\ell_{m_0}} g(F, m_{01}). \tag{2.4}$$

Because the optimization set is smaller than in (2.3), $\underline{\theta}_{0,\ell}$ and $\overline{\theta}_{0,\ell}$ are only inner bounds in general, namely $\underline{\theta}_{0,\ell} \geq \underline{\theta}_0$ and $\overline{\theta}_{0,\ell} \leq \overline{\theta}_0$. Caratheodory's result ensures however that under Assumption 2.1, these inner bounds coincide with the sharp bounds for $\ell = K+1$. In concrete terms, this means that for finding the sharp lower and upper bounds on the segregation index, we can make as if there was a finite number of types of firms with the same underlying probability.

[9] It suffices to choose $\nu(u,v) = u/[2v(1-v)]$ and $h(x, m_{01}) = |x - m_{01}|$ for the Duncan, $\nu(u,v) = 1 - u/[v\ln(v)]$ and $h(x, m_{01}) = x\ln(x)$ for the Theil and $\nu(u,v) = 1 - v^{-b/(1-b)}(1-v)^{-1}u^{1/(1-b)}$ and $h(x, m_{01}) = x^b(1-x)^{1-b}$ for the Atkinson index.

[10] Linearity of $F \mapsto \int h(x, m_{01})dF(x)$, together with convexity of $\mathcal{D}_{m_0}$ and continuity of $\nu(., m_{01})$, also implies that the identification region is the interval $[\underline{\theta}_0, \overline{\theta}_0]$.

Theorem 2.1 summarizes our discussion on the two cases.

Theorem 2.1 - *If* $m_0 \in \partial\mathcal{M}$, $\underline{\theta}_0 = \overline{\theta}_0 = g(F^*, m_{01})$, *where* F^* *is the cdf of a discrete distribution with at most* $L+1$ *support points, where* L *is the integer part of* $(K+1)/2$.
- *If* $m_0 \in \overset{\circ}{\mathcal{M}}$ *and Assumption 2.1 holds,* $\underline{\theta}_{0,K+1} = \underline{\theta}_0$ *and* $\overline{\theta}_{0,K+1} = \overline{\theta}_0$.

Note that $\mathcal{D}_{m_0}^{K+1}$ can be seen as a subset of $[0,1]^{2K+1}$, as any $F \in \mathcal{D}_{m_0}^{K+1}$ is defined by its support points and associated probabilities. As a result, $\overline{\theta}_{0,K+1}$ and $\underline{\theta}_{0,K+1}$ can be obtained as an optimization over a subset of $[0,1]^{2K+1}$. Noteworthy, the result would also apply to the lower bound of concave functionals of F. Because $g(., m_{01})$ is concave in the case of the Gini index, $\underline{\theta}_{0,K+1} = \underline{\theta}_0$ for the Gini. However, the upper bound cannot be obtained similarly.

The second result of Theorem 2.1 can be easily generalized to moment problems of the kind

$$\inf_{F \in \mathcal{D}} \int q(x) dF(x) \; s.t. \; \int r(x) dF(x) = 0, \tag{2.5}$$

where $q(x) \in \mathbb{R}$ while $r(x)$ belongs to $\mathbb{R}^K$. Here as well, the infimum is attained by distributions with at most $K+1$ support points, which makes the optimization feasible in practice. An example where bounds of an identification region satisfies Problem (2.5) is average marginal effects in binary choice panel data (See Lemma 7 of Chernozhukov et al., 2013, for such a result). In that case, F represents the distribution of fixed effects and the constraints correspond to the fact that the probabilities of all possible sequences of choices should match those of the data.

2.3 Additional results in special cases

In the interior case, computing the bounds still require a nonlinear optimization under constraints that are also nonlinear in the support points. Interestingly, the Chebyshev-Markov problem has been further simplified under additional assumptions, using the theory of Chebyshev systems (see, e.g. Krein and Nudel'man, 1977). More precisely, we consider the following condition.

Assumption 2.2 g *satisfies Assumption 2.1. Moreover* h *does not depend on* m_{01}, *is* $\mathcal{C}^{K+1}$ *on* $(0,1)$ *and satisfies either* $h^{(K+1)}(x) > 0$ *for all* $x \in (0,1)$ *or* $h^{(K+1)}(x) < 0$ *for all* $x \in (0,1)$.

Assumption 2.2 is satisfied for the Theil index. In this case, $h(x) = x \ln x$. h is C^∞ on $(0,1)$ and satisfies $h^{(K+1)}(x) = (-1)^{K+1}(K-1)!/x^K$ for $K \geq 1$. Thus $h^{(K+1)}$ has constant sign for all $K \geq 1$. In the case of the Atkinson index, $h(x) = x^b(1-x)^{1-b}$, we checked numerically that for all $b \in (0,1)$ and K odd between 3 and 49, $h^{K+1}(x) < 0$ for all $x \in (0,1)$, so that Assumption 2.2 is also satisfied for the Atkinson index for all odd $K \leq 50$.

Under Assumption 2.2, no numerical optimization is needed to compute the bounds $\underline{\theta}_0$ and $\overline{\theta}_0$. The idea behind is that special discrete distributions rationalizing the bounds, called principal representations, will also rationalize the bounds with $h(x) = x^{K+1}$.[11] Then, one can show that in the latter case, the problem reduces to finding the roots of a polynomial, a task for which very efficient algorithms are available. Using principal representations to compute the bounds is therefore much simpler and faster than solving (2.4), a point that we confirm below in our simulations (see in particular Table 3).

Let us now detail how the principal representations can be obtained. We do not provide proofs of our claims hereafter but refer to the monograph of Krein and Nudel'man (1977) for more details. The principal representations are determined solely by the vector $x = (x_1, ..., x_{L+1})$ of their support points, with $0 \leq x_1 < ... < x_{L+1} \leq 1$.[12] Then the associated vector of probabilities $y = (y_1, ..., y_{L+1})$ is uniquely defined by the $L+1$ moment constraints $V(x)y' = (1, E(p), ..., E(p^L))'$, where $V(x)$ is the Vandermonde matrix associated with vector x:[13]

$$V(x) = \begin{pmatrix} x_1^0 & \dots & x_{L+1}^0 \\ \vdots & & \\ x_1^L & \dots & x_{L+1}^L \end{pmatrix}.$$

y is uniquely defined by $V(x)y' = (1, m_0')'$ because Vandermonde matrices are nonsingular (see, e.g. Horn and Johnson, 1990).

Now, let us define the support points of the principal representations. For that

[11] Interestingly, principal representations have found numerous other applications in statistics, see Dette and Studden (1997) for a survey or Dette and Schorning (2013) for a recent application to optimal design of experiments.

[12] We consider here the case where the principal representations have $L+1$ support points. They may have less support points, in which case we should modify the dimension of x accordingly.

[13] In the following, Vandermonde matrices of different sizes will be used, depending on the size of x. In the absence of ambiguity, we keep the notations $V(x)$.

purpose, let A_{m_0}, B_{m_0} and C_{m_0} denote the square matrices of size L, L and $L-1$ respectively, with typical (i,j) term equal to m_{0i+j-2}, m_{0i+j-1} and $m_{0i+j}-m_{0i+j-1}$ respectively, with the convention that $m_{00}=1$. If K is even, first, define $\underline{a}=(\underline{a}_0,...,\underline{a}_{L-1})'$ and $\overline{a}=(\overline{a}_0,...,\overline{a}_{L-1})'$ by

$$\begin{aligned}\underline{a} &= -B_{m_0}^{-1}\left(m_{0L+1},...,m_{0K}\right)', \\ \overline{a} &= \left(B_{m_0}-A_{m_0}\right)^{-1}\left(m_{0L}-m_{0L+1},...,m_{0K-1}-m_{0K}\right)'.\end{aligned} \tag{2.6}$$

That B_{m_0} and $B_{m_0}-A_{m_0}$ are nonsingular is ensured by $m_0\in\overset{\circ}{\mathcal{M}}$ and K even (see Remark III 2.1 of Krein and Nudel'man, 1977). Then consider the polynomials $\underline{P}_{m_0}$ and $\overline{P}_{m_0}$ defined by

$$\underline{P}_{m_0}(x)=\sum_{j=0}^{L-1}\underline{a}_j x^j+x^L,\ \overline{P}_{m_0}(x)=\sum_{j=0}^{L-1}\overline{a}_j x^j+x^L.$$

The subscript m_0 underlines the dependency of these polynomials on m_0, through (2.6). The support points of the lower principal representation $\underline{F}_{m_0}$ are then 0 and the roots of $\underline{P}_{m_0}$. Similarly, the support points of the upper principal representation $\overline{F}_{m_0}$ are 1 and the roots of $\overline{P}_{m_0}$. The construction is the same in the odd case. $\underline{a}$ and $\overline{a}=(\overline{a}_0,...,\overline{a}_{L-2})'$ then satisfy

$$\underline{a}=-A_{m_0}^{-1}\left(m_{0L},...,m_{0K}\right)',\ \overline{a}=C_{m_0}^{-1}\left(m_{0L}-m_{0L+1},...,m_{0K-1}-m_{0K}\right)'.$$

The polynomials $\underline{P}_{m_0}$ and $\overline{P}_{m_0}$ are defined similarly, and the support points of $\underline{F}_{m_0}$ (resp. $\overline{F}_{m_0}$) are the roots of $\underline{P}_{m_0}$ (resp. 0, 1 and the roots of $\overline{P}_{m_0}$).

In the case of the Atkinson index, Assumption 2.2 does not hold for K even. In this case, however, we can still rely on Chebyshev systems, by remarking that $h(x)=x^{K+1}$ satisfies Assumption 2.2. In other words, the lower and upper bounds on m_{0K+1}, denoted respectively by $\underline{m}_{0K+1}$ and $\overline{m}_{0K+1}$, can be obtained by the previous construction. Then one possibility would be to compute the bounds on A_b given $(m_{01},...,m_{0K+1})$, for all possible values of m_{0K+1} in $[\underline{m}_{0K+1},\overline{m}_{0K+1}]$. But the bounds on A_b are even simpler to yield, by properties of Chebyshev systems. Specifically, Theorem VI.2.2 of Krein and Nudel'man (1977) ensures that the bounds on A_b are attained on either $\underline{m}_{0K+1}$ or $\overline{m}_{0K+1}$.

We summarize our discussion in the following theorem.

Theorem 2.2 *Suppose that Assumption 2.2 holds. Then*

$$\left\{\underline{\theta}_0,\overline{\theta}_0\right\}=\left\{g(\underline{F}_{m_0},m_{01}),g(\overline{F}_{m_0},m_{01})\right\}.$$

Moreover, if Assumption 2.2 holds for $K+1$ instead of K, then

$$\underline{\theta}_0 = \min\left\{g\left(\underline{F}_{\underline{m}_0}, m_{01}\right), g\left(\underline{F}_{\overline{m}_0}, m_{01}\right), g\left(\overline{F}_{\underline{m}_0}, m_{01}\right), g\left(\overline{F}_{\overline{m}_0}, m_{01}\right)\right\},$$
$$\overline{\theta}_0 = \max\left\{g\left(\underline{F}_{\underline{m}_0}, m_{01}\right), g\left(\underline{F}_{\overline{m}_0}, m_{01}\right), g\left(\overline{F}_{\underline{m}_0}, m_{01}\right), g\left(\overline{F}_{\overline{m}_0}, m_{01}\right)\right\},$$

where $\underline{m}_0 = (m_{01}, ..., m_{0K}, \underline{m}_{0K+1})$ and $\overline{m}_0 = (m_{01}, ..., m_{0K}, \overline{m}_{0K+1})$.

2.4 Links with other approaches

Previous approaches in the literature have focused 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 $\theta_N = g(F_{X/K}, m_{01})$. However, the following proposition shows that if $g(., m_{01})$ is monotonic with respect to the second-order dominance, as is the case of all the inequality indices we consider, this parameter is always greater than $\overline{\theta}_0$. In other words, ignoring the randomness leads to overestimate the true level of segregation.

Proposition 2.3 *Suppose that $g(., m_{01})$ is decreasing with respect to the second-order dominance. Then $\theta_N \geq \overline{\theta}_0$. Moreover, the inequality is strict if $g(., m_{01})$ is strictly decreasing*[14] *and the support of p is not reduced to $\{0, 1\}$.*

Several works have recognized this small-unit bias. The most commonly used correction method is the one introduced by Carrington and Troske (1997), based on earlier works by Winship (1977) and Cortese et al. (1978). The idea is to define an index that corresponds to a distance from randomness. Specifically, let $\theta_N^{ns} = g(F_{X^{ns}/K}, m_{01})$, with $X^{ns} \sim B(E(p), K)$, denote the naive parameter that would be obtained if all units had the same probability, that is if there was no segregation. Suppose also, without loss of generality if g is bounded, that g ranges from 0 to 1. The corrected index θ_{CT} of Carrington and Troske (1997) is defined by

$$\theta_{CT} = \frac{\theta_N - \theta_N^{ns}}{1 - \theta_N^{ns}}.$$

The index θ_{CT} is therefore an affine correction that coincides with θ_0 in the two polar cases where there is no segregation, because $\theta_{CT} = \theta_N = \theta_0 = 0$ in this case,

[14]Here we say that $g(., m_{01})$ is strictly decreasing with respect to the second-order dominance if, whenever $\int w(x)dF(x) > \int w(x)dG(x)$ for all strictly concave w, we have $g(F, m_{01}) > g(G, m_{01})$.

or if segregation is maximal, because then $\theta_{CT} = \theta_0 = 1$. But in in general θ_{CT} is not equal to θ_0, nor does it lie inside the interval $[\underline{\theta}_0, \overline{\theta}_0]$, as we will illustrate in Subsection 4.1.

Allen et al. (2015) propose a bootstrap correction of the segregation index. Their method aims to obtain a good approximation of the discrepancy between $\theta_N = g(F_{X/K}, m_{01})$ and θ_0 by bootstrap, and then to correct for this discrepancy. In our framework, this would amount to approximate this discrepancy by $\theta_N^* - \theta_N$, where $\theta_N^* = g(F_{X^*/K}, m_{01})$ and $X^*|X \sim B(K, X/K)$.[15] The corrected index is then:

$$\theta_{ABW} = 2\theta_N - \theta_N^* \ (= \theta_N + \theta_N - \theta_N^*).$$

The idea behind this parameter is that, if X/K was distributed as p, we would have $\theta_N - \theta_0 = \theta_N^* - \theta_N$ and $\theta_{ABW} = \theta_0$. More generally, one can show that the bias of θ_{ABW} decreases more quickly than the one of θ_N as $K \to \infty$.

If focusing on θ_0 rather than θ_{ABW} or θ_{CT} raises some identification issues, an important advantage of our approach is that it sticks to indices whose axiomatic properties are well understood (see, e.g., James and Taeuber, 1985; Chakravarty and Silber, 1994; Hutchens, 2001). One particularly desirable property is size invariance, satisfied by all indices we consider (James and Taeuber, 1985). While θ_{ABW} or θ_{CT} correct for part of the small-unit bias, the resulting index will in general depend on the unit size and violate the size invariance principle.

Rathelot (2012) follows a closer approach to ours by considering the same parameter θ_0. But contrary to us, he imposes the distribution of p to be a mixture of beta distributions. Combined with the binomial assumption on X, the model becomes fully parametric and can be estimated by maximum likelihood. The segregation indices can be easily deduced as a function of the parameters of the beta mixture. Note that such a model is overidentified in general. For instance, a mixture of two beta distributions has five parameters, so that most vectors of first K moments will not be compatible with this model when $K \geq 6$. In such cases, the segregation index obtained may not lie inside the interval $[\underline{\theta}_0, \overline{\theta}_0]$. Importantly, this corrected index will only converge to θ_0 as $K \to \infty$ if one lets the number of components of the mixture tend to infinity with K.

[15] In their framework, θ_N^* is not exactly defined this way, because the two allocation models differ. The two are however expected to be close when the sample size is large, for the reasons detailed in Footnote 6.

3 Estimation and inference

3.1 Estimation of the bounds

In this section, we suppose to have in hand an i.i.d. sample $(X_1, ..., X_n)$ of n units. Unit sizes are still constant equal to K. Following the identification part, we estimate the identified set by estimating its sharp bounds. We first estimate P_0 and thus $m_0 = Q^{-1}P_0$. We then use Theorems 2.1 or 2.2 to yield the estimates of the bounds.

First, we estimate P_0 by constrained maximum likelihood, where the constraints come from the binomial mixture model. By what precedes, the model is equivalent to $P_0 \in \mathcal{P} = \{Qm : m \in \mathcal{M}\}$. We then let

$$\widehat{P} = \arg\max_{P \in \mathcal{P}} \sum_{k=1}^{K} N_k \ln(P_k) + N_0 \ln\left(1 - \sum_{k=1}^{K} P_k\right), \tag{3.1}$$

where $N_k = \sum_{i=1}^{n} \mathbb{1}\{X_i = k\}$. This optimization may look complicated because $\mathcal{M}$, and thus $\mathcal{P}$, is defined in a complicated way. We can use however a simpler characterization of $\mathcal{M}$ to simplify it much, as Lemma 3.1 below shows. Hereafter, we let $\mathcal{S}_{L+1} = \{(x_1, ..., x_{L+1}) : 0 \leq x_1 < ... < x_{L+1} \leq 1\}$ and $\mathcal{T}_{L+1} = \{(y_1, ..., y_{L+1}) \in [0,1]^{L+1} : \sum_{k=1}^{L+1} y_k = 1\}$.

Lemma 3.1 *The maximum likelihood estimator $\widehat{P} = (\widehat{P}_1, ..., \widehat{P}_K)'$ satisfies*

$$\widehat{P}_k = \binom{K}{k} \sum_{j=1}^{L+1} \widehat{y}_j \widehat{x}_j^k (1 - \widehat{x}_j)^{K-k}, \quad k \in \{1, ..., K\},$$

where $\widehat{x} = (\widehat{x}_1, ..., \widehat{x}_{L+1})$ and $\widehat{y} = (\widehat{y}_1, ..., \widehat{y}_{L+1})$ are given by

$$(\widehat{x}, \widehat{y}) = \arg \max_{(x,y) \in \mathcal{S}_{L+1} \times \mathcal{T}_{L+1}} \sum_{k=0}^{K} N_k \ln\left(\sum_{j=1}^{L+1} y_j x_j^k (1 - x_j)^{K-k}\right).$$

Following (2.2), we then estimate m_0 by $\widehat{m} = Q^{-1}\widehat{P}$. Note that by construction, $\widehat{m} \in \mathcal{M}$.

Now let us turn to the segregation index. We rely on Theorems 2.1 and 2.2 to estimate its bounds. We first check whether $\widehat{m} \in \partial\mathcal{M}$ or not, because if this is the case, the bounds are equal and no optimization is required. A simple way to test this is to consider whether the unconstrained maximum likelihood estimator $\widetilde{P} = (\widetilde{P}_1, ..., \widetilde{P}_K)$, which satisfies $\widetilde{P}_k = N_k/n$, belongs or not to $\mathcal{P}$. We propose a

simple procedure for testing $\widetilde{P} \notin \mathcal{P}$ in Subsection D.1 of the online appendix. Our Monte Carlo simulations show that $\widehat{m} \in \partial\mathcal{M}$ occurs with probability close to one when $K \geq 10$, even with sample sizes as large as 10,000 (for similar evidence, see Wood, 1999). In this case, we simply estimate the bounds by

$$\widehat{\underline{\theta}} = \widehat{\overline{\theta}} = g(\widehat{F}, \widehat{m_1}), \tag{3.2}$$

where $\widehat{F}$ is the cdf corresponding to $(\widehat{x}, \widehat{y})$.

If $\widetilde{P} \in \mathcal{P}$, $\widehat{m} \in \overset{\circ}{\mathcal{M}}$ with probability approaching one.[16] Then $\mathcal{D}_{\widehat{m}}$ is not reduced to a single distribution, and if Assumption 2.2 is not satisfied, optimization is required to obtain the estimated bounds. We then use estimators of $\overline{\theta}_{0,K+1}$ and $\underline{\theta}_{0,K+1}$. Given a vector of moments $m = (m_1, ..., m_K)$, any $F \in \mathcal{D}_m^{K+1}$ is defined by its support points $x \in \mathcal{S}_{K+1}$ and the associated probabilities $y \in \mathcal{T}_{K+1}$. Moreover, the moment constraints write $V(x)y = (1, m')'$. Thus, the vector of probabilities y satisfies $y' = V(x)^{-1}(1, m')'$, and the constraints are equivalent to $V(x)^{-1}(1, m')' \geq 0$, where the inequalities are understood componentwise. Because $F \in \mathcal{D}_m^{K+1}$ depends on x and m only, we may rewrite $g(F, m_1)$ as a function of x and m only. We denote this function by $q(x, m)$. The bounds on the true parameter $\theta_0 = \theta(m)$ when the vector of moments is $m_0 = m$ satisfy

$$\underline{\theta}(m) = \min_{x \in \mathcal{S}_{K+1} : V(x)^{-1}(1,m')' \geq 0} q(x, m), \tag{3.3}$$

$$\overline{\theta}(m) = \max_{x \in \mathcal{S}_{K+1} : V(x)^{-1}(1,m')' \geq 0} q(x, m). \tag{3.4}$$

Our estimators of $\overline{\theta}_0$ and $\underline{\theta}_0$ are respectively $\widehat{\underline{\theta}} = \underline{\theta}(\widehat{m})$ and $\widehat{\overline{\theta}} = \overline{\theta}(\widehat{m})$.

Finally, when Assumption 2.2 holds, we simply estimate the principal representations $\underline{F}_{m_0}$ and $\overline{F}_{m_0}$ by $\underline{F}_{\widehat{m}}$ and $\overline{F}_{\widehat{m}}$, and let

$$\widehat{\underline{\theta}} = \min\left\{g\left(\underline{F}_{\widehat{m}}, \widehat{m_1}\right), g\left(\overline{F}_{\widehat{m}}, \widehat{m_1}\right)\right\}, \quad \widehat{\overline{\theta}} = \max\left\{g\left(\underline{F}_{\widehat{m}}, \widehat{m_1}\right), g\left(\overline{F}_{\widehat{m}}, \widehat{m_1}\right)\right\}. \tag{3.5}$$

3.2 Inference on the segregation index and its identified set

We first show that the estimators of the bounds are root-n consistent and characterize their asymptotic distribution. We consider hereafter both the cases where $m_0 \in \overset{\circ}{\mathcal{M}}$ and $m_0 \in \partial\mathcal{M}$, since the corresponding asymptotic distributions differ. We obtain the result under the following two conditions.

[16] The only exception is when $\widetilde{P} \in \partial\mathcal{P}$, the boundary of $\mathcal{P}$. This occurs however with probability tending to 0 as $n \to \infty$.

Assumption 3.1 *The distribution of p is not a Bernoulli distribution.*

Assumption 3.2 *$\underline{\theta}$ and $\overline{\theta}$ are directionally differentiable at m_0 in the following sense: $\underline{\theta}'(m,h) = \lim_{t\downarrow 0}(\underline{\theta}(m+th_t) - \underline{\theta}(m))/t$ exists for all $h_t \in \mathbb{R}^K$ such that $h_t \to h$ and $m + th_t \in \mathcal{M}$ for t small enough. Moreover, $\underline{\theta}'(m,.)$ is continuous.*

The first assumption excludes total segregation, where we would either have units with only people from the minority group or only people from the majority. We rule out such situations for inference, because estimators are then degenerated, namely they coincide with the true values. Assumption 3.2 is more substantial, but can be proved to hold in two cases of interest (see Subsection D.2 of the online appendix).

Before giving the asymptotic distribution of the estimated bounds, we introduce additional notations. For any vector P, let us define $\Sigma(P) = [\text{diag}(P) - PP']$, $\text{diag}(P)$ being the diagonal matrix with diagonal vector equal to P. We let $C_{P_0} = \{\lambda(P - P_0),\ P \in \mathcal{P}, \lambda > 0\}$ and $\pi_{\overline{C}_{P_0}}$ the projection onto the closure of C_{P_0} with respect to the norm $\|x\| = x' \left(\text{diag}(P_0)^{-1} + M_1/P_{00}\right) x$, M_1 being the $K \times K$ matrix of ones.

Theorem 3.1 *Suppose that Assumption 2.1 holds. Then $(\widehat{\underline{\theta}}, \widehat{\overline{\theta}}) \xrightarrow{P} (\underline{\theta}, \overline{\theta})$. If Assumptions 3.1-3.2 also hold, then*

$$\sqrt{n}\left(\widehat{\underline{\theta}} - \underline{\theta}_0, \widehat{\overline{\theta}} - \overline{\theta}_0\right)' \xrightarrow{d} \left(\underline{\theta}'\left(m_0, Q^{-1}\pi_{\overline{C}_{P_0}}(\mathcal{Z})\right), \overline{\theta}'\left(m_0, Q^{-1}\pi_{\overline{C}_{P_0}}(\mathcal{Z})\right)\right)',$$

where $\mathcal{Z} \sim \mathcal{N}(0, \Sigma(P_0))$.

Importantly, our results apply whether or not m_0 lies in the interior of $\mathcal{M}$. If $m_0 \in \overset{\circ}{\mathcal{M}}$ and $\underline{\theta}$ and $\overline{\theta}$ are differentiable rather than simply directionally differentiable, the estimated bounds are asymptotically normal, because $\pi_{\overline{C}_{P_0}}(\mathcal{Z}) = \mathcal{Z}$. But if $m_0 \in \partial\mathcal{M}$, the asymptotic distribution of the estimated bounds is a function of the projection of a normal variable onto a convex cone.

Because the estimated bounds are not asymptotically normal when $m_0 \in \partial\mathcal{M}$, the confidence interval proposed by Imbens and Manski (2004) for partially identified parameters does not apply here. Moreover, standard bootstrap typically fails here, because of the lack of continuity in m_0 of the asymptotic distribution (see Andrews, 2000, for a similar counterexample). To build valid confidence intervals, we therefore propose a modified bootstrap procedure. We project $\widehat{m}$ onto $\partial\mathcal{M}$

whenever $\widehat{m} \in \overset{\circ}{\mathcal{M}}$ but is close to the boundary. Let $d_n = \sqrt{n}(\widehat{\overline{\theta}} - \widehat{\underline{\theta}})/k_n$ and $I_n = \mathbb{1}\{d_n \leq 1\}$, with $k_n \to \infty$, $\sqrt{n}/k_n \to \infty$. Observe that when $m_0 \in \partial\mathcal{M}$, $\underline{\theta} = \overline{\theta}$ so that $d_n \xrightarrow{P} 0$ and $I_n \xrightarrow{P} 1$. When $m_0 \in \overset{\circ}{\mathcal{M}}$ on the other hand, $\underline{\theta} < \overline{\theta}$ in general because there is an infinity of distributions rationalizing m_0. Thus $d_n \xrightarrow{P} \infty$ and $I_n \xrightarrow{P} 0$. Then we define

$$\widehat{m}_b = \pi_{\partial\mathcal{M}}(\widehat{m})I_n + \widehat{m}(1 - I_n),$$

where $\pi_{\partial\mathcal{M}}$ denotes the projection onto $\partial\mathcal{M}$.[17] The bootstrap distribution of X that we consider hereafter is given by the vector of probabilities $\widehat{P}_b = Q\widehat{m}_b$.

We now define the bootstrap confidence intervals. We have to take into account the fact that the lower and upper bounds collapse when $m_0 \in \partial\mathcal{M}$, whereas they are in general distinct when $m_0 \in \overset{\circ}{\mathcal{M}}$. For any statistic T, let T^* denote the corresponding bootstrap statistic. For example, if $\underline{T} = \sqrt{n}(\widehat{\underline{\theta}} - \underline{\theta})$, we let $\underline{T}^* = \sqrt{n}(\widehat{\underline{\theta}}^* - \widehat{\underline{\theta}})$, where $\widehat{\underline{\theta}}^*$ is the bootstrap estimator of $\underline{\theta}$. We let $c_\alpha(T^*)$ denote the α-th quantile of the distribution of T^* conditional on $\widehat{m}_b$. We first define a confidence interval for the interior case by

$$\mathrm{CI}_{1-\alpha}^{\text{interior}} = \left[\widehat{\underline{\theta}} - \frac{c_{1-\alpha}(\underline{T}^*)}{\sqrt{n}},\ \widehat{\overline{\theta}} - \frac{c_\alpha(\overline{T}^*)}{\sqrt{n}}\right],$$

where $\overline{T}$ is defined as $\underline{T}$. The reason why we use $c_\alpha(\overline{T}^*)$ and $c_{1-\alpha}(\underline{T}^*)$ instead of $c_{\alpha/2}(\overline{T}^*)$ and $c_{1-\alpha/2}(\underline{T}^*)$ is that when $m_0 \in \overset{\circ}{\mathcal{M}}$, $\underline{\theta}_0 < \overline{\theta}_0$ in general and only one of the two bounds matter in the asymptotic coverage.

This is not the case however when $m_0 \in \partial\mathcal{M}$. Because $\underline{\theta}_0 = \overline{\theta}_0 = \theta_0$, the asymptotic coverage of $\mathrm{CI}_{1-\alpha}^{\text{interior}}$ is in general smaller than $1 - \alpha$. We consider instead the symmetric confidence interval

$$\mathrm{CI}_{1-\alpha}^{\text{boundary}} = \left[\widehat{\theta} - \frac{c_{1-\alpha}(T_s^*)}{\sqrt{n}},\ \widehat{\theta} + \frac{c_{1-\alpha}(T_s^*)}{\sqrt{n}}\right],$$

where $\widehat{\theta} = (\widehat{\underline{\theta}} + \widehat{\overline{\theta}})/2$ and $T_s = \sqrt{n}\left|\widehat{\theta} - (\underline{\theta}_0 + \overline{\theta}_0)/2\right|$. When $m_0 \in \partial\mathcal{M}$, $\widehat{\theta}$ is a consistent estimator of $\theta_0 = (\underline{\theta}_0 + \overline{\theta}_0)/2$ and we show in the proof of Theorem 3.2 below that the bootstrap statistic T_s^* has the same distribution as T_s. Thus, $\mathrm{CI}_{1-\alpha}^{\text{boundary}}$ has an asymptotic coverage rate of $1 - \alpha$.

[17] Because $\partial\mathcal{M}$ is not convex, this projection may not be well defined. This is not an issue here. In this case, $\pi_{\partial\mathcal{M}}(\widehat{m})$ denotes any element in the set $\arg\min_{m \in \partial\mathcal{M}} \|\widehat{m} - m\|$.

Finally, to obtain a confidence interval with a correct asymptotic coverage in all situations, we let

$$\text{CI}^1_{1-\alpha} = I_n \text{CI}^{\text{boundary}}_{1-\alpha} + (1 - I_n)\text{CI}^{\text{interior}}_{1-\alpha}.$$

The idea is that we will eventually pick $\text{CI}^{\text{boundary}}_{1-\alpha}$ when the true parameter is at the boundary, because $I_n \xrightarrow{P} 1$ in this case, and $\text{CI}^{\text{interior}}_{1-\alpha}$ otherwise. The validity of this confidence interval, established in Theorem 3.2 below, relies on the following condition.

Assumption 3.3 *$\underline{\theta}(.)$ and $\overline{\theta}(.)$ are differentiable at m_0 (and we let $\underline{\theta}'(m_0)$ and $\overline{\theta}'(m_0)$ denote their gradient). Moreover, we either have*

- *$m_0 \in \overset{\circ}{\mathcal{M}}$, $\underline{\theta}_0 < \overline{\theta}_0$ and $\overline{\theta}'(m_0) \neq 0$, $\underline{\theta}'(m_0) \neq 0$;*
- *or $m_0 \in \partial\mathcal{M}$, with $\overline{C}_{m_0}$ a half space and the cdf of the asymptotic distribution of T_s continuous at its $1-\alpha$ quantile.*

Assumption 3.3 is rather mild. Lemma D.2 already shows that the bounds are differentiable almost everywhere in several cases. When $m_0 \in \overset{\circ}{\mathcal{M}}$, the set of distributions $\mathcal{D}_{m_0}$ is infinite, so that $\underline{\theta}_0 < \overline{\theta}_0$ holds in general. The important restriction, when $m_0 \in \partial\mathcal{M}$, is that $\mathcal{M}$ is smooth at m_0, so that $\overline{C}_{m_0}$ is a half space. This holds everywhere except at $(0,0)$ and $(1,1)$ when $K = 2$, because in this case $\partial\mathcal{M} = \{(m_{01}, m_{01}), m_{01} \in [0,1]\} \cup \{(m_{01}, m_{01}^2), m_{01} \in [0,1]\}$. We conjecture that it also holds almost everywhere when $K \geq 3$, though the analysis of the geometry of $\partial\mathcal{M}$ is beyond the scope of the paper.

Theorem 3.2 *Suppose that Assumptions 2.1, 3.1, 3.2 and 3.3 hold. Then, with probability one,*

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

Theorem 3.2 shows that bootstrap confidence intervals are asymptotically valid in general. The conditions for obtaining this result are the differentiability of the bounds and the fact when $m_0 \in \partial\mathcal{M}$, $\overline{C}_{m_0}$ is a half space. Theoretically speaking, it is possible to drop these conditions and still make valid inference by using subsampling, as for instance Chernozhukov et al. (2007) or Romano and Shaikh (2010). However, Monte Carlo simulations (not reported here) seem to

indicate that, in our context, subsampling does not provide reliable results unless the sample size n is very large.

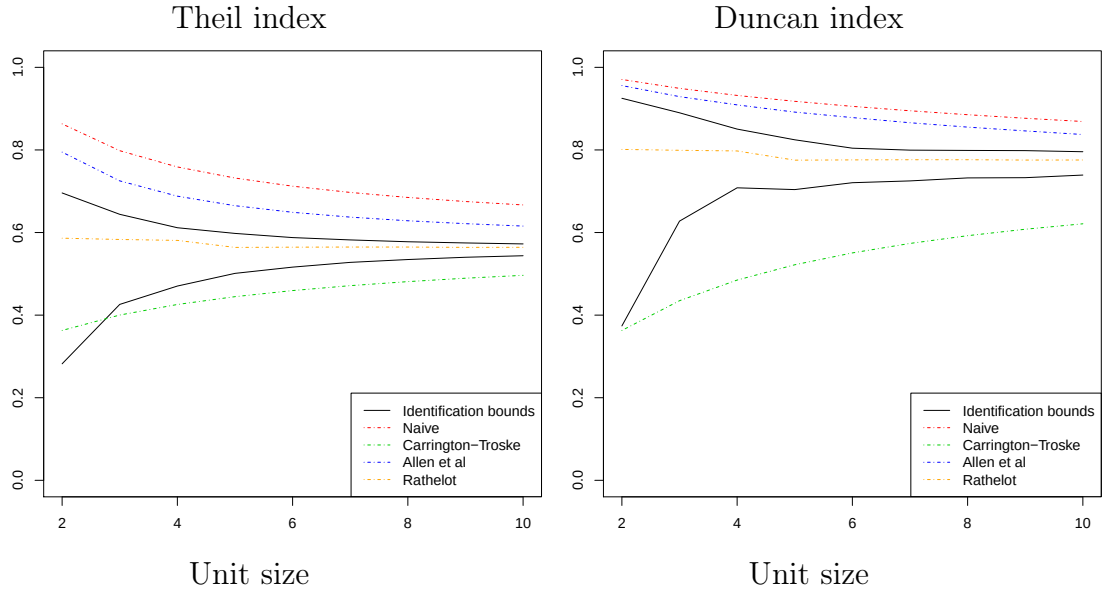
$\text{CI}^1_{1-\alpha}$ is asymptotically valid whether m_0 lies in the interior or at the boundary of $\mathcal{M}$. It is unclear, on the other hand, whether it is uniformly valid. The confidence interval considered by Imbens and Manski (2004) in a related setting is uniformly valid, but this is because they assume a uniform convergence in distribution of the estimated bounds. Such a uniform convergence does not hold here, as asymptotic normality fails to hold at the boundary. That inference on a partially identified parameter may not be uniform is underlined by Andrews and Han (2009), in a related context where the endpoints of the identification interval are estimated. We consider in the online appendix (see Subsection D.3) another confidence interval that satisfies the uniformity requirement but is generally conservative.

4 Simulations

4.1 Identified bounds and other approaches

Figure 1 presents a comparison, for the Theil and Duncan index, between the sharp bounds, the naive approach and the corrections proposed by Carrington and Troske (1997), Allen et al. (2015) and Rathelot (2012). We consider $\Phi^{-1}(p) \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu \simeq -3.12$ and $\sigma^2 \simeq 1.56$ chosen so as to be close to the first two estimated moments of p in our application in Section 5. The sharp bounds are obtained by solving (2.4), the naive and the Carrington and Troske parameter by using their theoretical expressions, the Allen et al. corrected index by simulations on a very large sample ($n = 10^6$) and the corrected index of Rathelot (2012) by maximizing the theoretical log-likelihood of the model.

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



Note: $\Phi^{-1}(p) \sim \mathcal{N}(-3.12, 1.56)$. With this DGP, the Theil index is $T \simeq 0.562$ and the Duncan index is $D \simeq 0.775$.

Firstly, the length of the identification region shrinks quickly between $K = 2$ and $K = 6$ for both indices, less so after. As expected, the naive approach is well above the upper bound of the identification region. For both indices, the corrected indices proposed by Allen et al. (2015) or Carrington and Troske (1997) always lie outside of the identification interval: the former is always above and the latter always below (except for the Theil index with $K = 2$). The correction proposed by Carrington and Troske (1997) performs better with the Theil than with the Duncan index. The parametric method of Rathelot (2012) lies within the bound for all $K \leq 10$ with this DGP, but this needs not be the case in general.

Table 1 presents a comparison of the different approaches when the unit size is random, and uniform on $\{2, ..., 10\}$. As discussed in Subsection B.2 of the online appendix, we can consider an "unweighted" index, focused on the unit, (Equation (B.1)) or a "weighted" index, focused on the worker (Equation (B.2)). With our DGP for which $K \perp\!\!\!\perp p$, the two indices coincide but they lead to different identification sets, because the identification interval shrinks with K and larger values of K are weighted more with the individual-weighted index. Consistent with the results obtained with a fixed unit size, the naive index as well as the corrected indices by Carrington and Troske or Allen et al. do not lie in the identification

set. Conversely, the corrected index by Rathelot does, in this case.

Table 1: Comparison between the sharp bounds, the naive approach and previous corrections, with a random unit size.

Method	Theil index	Duncan index
Sharp bounds		
Unweighted	[0.48,0.60]	[0.69,0.83]
Weighted	[0.51,0.59]	[0.72,0.81]
Naive	0.74	0.92
Carrington-Troske	0.45	0.55
Allen et al.	0.68	0.89
Rathelot	0.57	0.78

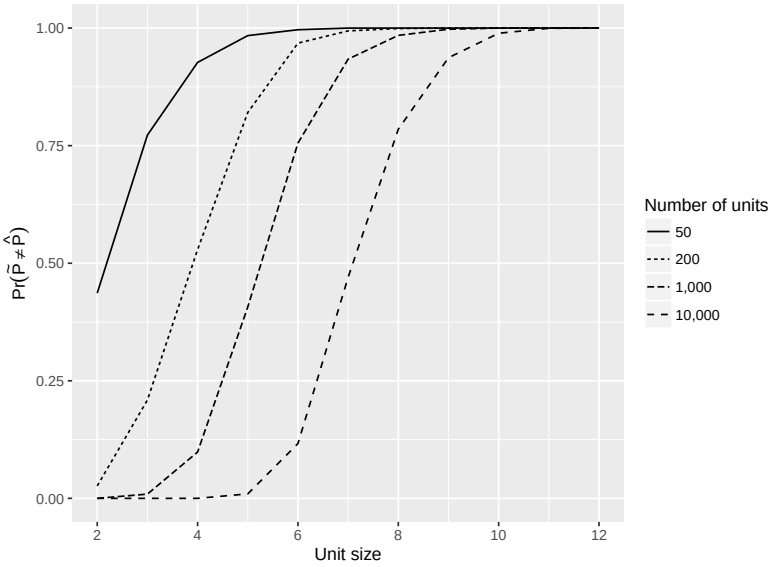
Note: $\Phi^{-1}(p) \sim \mathcal{N}(-3.12, 1.56)$. With this DGP, the Theil index is $T \simeq 0.562$ and the Duncan index is $D \simeq 0.775$.

4.2 Monte Carlo simulations

We now assess the performance of the estimators and confidence intervals considered in this paper in order to solve small-unit biases. We first study whether the constraint that P_0 belongs to $\mathcal{P}$ is binding in practice when estimating P_0. The data generating process is defined as previously ($\Phi^{-1}(p) \sim \mathcal{N}(-3.12,\ 1.56)$), and we estimate $\Pr(\widetilde{P} \notin \mathcal{P})$ for different sample and unit sizes. Figure 2 presents the results for $n \in \{50; 200; 1000; 10000\}$ 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 $\mathcal{P}$ shrinks very quickly with K. For instance, with 200 units, the estimated probability (with 1,000 simulations) is one as soon as K is 7. Obviously, the probability is systematically lower when n is larger 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 for the estimated identification region of θ_0, even though the true identification region is not reduced to a singleton. Hence, 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 $\widehat{P}$ is constrained ($\Pr(\widetilde{P} \notin \mathcal{P})$).



Note: each dot corresponds to 1,000 simulations with the DGP $\Phi^{-1}(p) \sim \mathcal{N}(-3.12, 1.56)$.

Table 2 displays the properties of the estimated bounds and the confidence intervals $\text{CI}^1_{0.95}$ for different sample sizes. We consider here both the Theil and Duncan indices, and the data generating process is defined as before by $\Phi^{-1}(p) \sim \mathcal{N}(-3.12, 1.56)$. For this distribution, $T \simeq 0.562$ and $D \simeq 0.775$. $\text{CR}(\theta_0)$ denotes

Table 2: Performance of $[\widehat{\underline{\theta}}, \widehat{\overline{\theta}}]$ and properties of $\mathrm{CI}^1_{0.95}$.

		Theil index			Duncan index		
K	n	$[\underline{\theta}_0, \overline{\theta}_0]$	$[E(\widehat{\underline{\theta}}), E(\widehat{\overline{\theta}})]$ $(\sigma(\widehat{\underline{\theta}}))$ $(\sigma(\widehat{\overline{\theta}}))$	$\mathrm{CR}(\theta_0)$	$[\underline{\theta}_0, \overline{\theta}_0]$	$[E(\widehat{\underline{\theta}}), E(\widehat{\overline{\theta}})]$ $(\sigma(\widehat{\underline{\theta}}))$ $(\sigma(\widehat{\overline{\theta}}))$	$\mathrm{CR}(\theta_0)$
3	100	[0.426, 0.644]	[0.494 (0.171), 0.593 (0.170)]	0.928	[0.627, 0.890]	[0.701 (0.213), 0.818 (0.184)]	0.960
	1,000		[0.434 (0.064), 0.639 (0.043)]	0.998		[0.635 (0.088), 0.881 (0.041)]	0.975
	10,000		[0.426 (0.019), 0.643 (0.013)]	1.000		[0.626 (0.027), 0.886 (0.030)]	0.988
6	100	[0.516, 0.588]	[0.536 (0.112), 0.537 (0.112)]	0.958	[0.721 0.804]	[0.770 (0.107), 0.770 (0.107)]	0.970
	1,000		[0.541 (0.053), 0.552 (0.052)]	0.975		[0.774 (0.056), 0.787 (0.046)]	0.950
	10,000		[0.524 (0.027), 0.579 (0.027)]	0.930		[0.738 (0.035), 0.801 (0.020)]	0.978
9	100	[0.540, 0.575]	[0.544 (0.092), 0.544 (0.092)]	0.955	[0.733, 0.798]	[0.772 (0.080), 0.772 (0.080)]	0.963
	1,000		[0.552 (0.037), 0.552 (0.037)]	0.958		[0.781 (0.034), 0.781 (0.034)]	0.952
	10,000		[0.555 (0.023), 0.557 (0.023)]	0.998		[0.779 (0.024), 0.782 (0.023)]	0.985
12	100	[0.549, 0.569]	[0.538 (0.092), 0.538 (0.092)]	0.920	[0.753, 0.788]	[0.769 (0.075), 0.769 (0.075)]	0.960
	1,000		[0.557 (0.032), 0.557 (0.032)]	0.955		[0.780 (0.028), 0.780 (0.028)]	0.975
	10,000		[0.557 (0.018), 0.557 (0.018)]	0.990		[0.775 (0.024), 0.775 (0.024)]	1.000
Random u.							
	100	[0.508, 0.594]	[0.547 (0.062), 0.563 (0.060)]	0.942	[0.708, 0.820]	[0.758 (0.063), 0.790 (0.059)]	0.884
	1,000		[0.526 (0.026), 0.577 (0.022)]	0.958		[0.742 (0.028), 0.808 (0.023)]	0.998
	10,000		[0.519 (0.011), 0.583 (0.011)]	1.000		[0.730 (0.013), 0.811 (0.010)]	1.000
Random w.							
	100	[0.528, 0.582]	[0.546 (0.052), 0.557 (0.052)]	0.947	[0.728, 0.804]	[0.768 (0.044), 0.781 (0.042)]	0.938
	1,000		[0.542 (0.022), 0.565 (0.021)]	0.894		[0.764 (0.021), 0.792 (0.019)]	0.924
	10,000		[0.537 (0.012), 0.569 (0.012)]	0.944		[0.754 (0.013), 0.793 (0.012)]	0.961

Note: for each (n, K), simulations are based on 400 draws of samples. The distribution of p is $\Phi^{-1}(p) \sim \mathcal{N}(-3.12, 1.56)$, leading to $T \simeq 0.562$ and $D \simeq 0.775$. $\mathrm{CR}(\theta_0) = \Pr(\theta_0 \in \mathrm{CI}^1_{0.95})$. "Random" corresponds to a random K, drawn with equal probability in $\{3, 6, 9, 12\}$. "u." and "w." refer respectively to the unweighted and weighted indices defined by (B.1) and (B.2).

the coverage rate of the true parameter by the confidence interval. We consider designs with fixed K in $\{3, 6, 9, 12\}$ as well as a random design where K is drawn in this same set with equal probability. Finally, to build confidence intervals, we use, following the law of iterated logarithm, $k_n = \left(2 \ln \ln(n)/[n\widehat{V}^*(\widehat{\overline{\theta}} - \widehat{\underline{\theta}})]\right)^{1/2} \mathbb{1}\{\widehat{\overline{\theta}} > \widehat{\underline{\theta}}\}$, where $\widehat{V}^*(\widehat{\overline{\theta}} - \widehat{\underline{\theta}})$ denotes the bootstrap estimator of $V(\widehat{\overline{\theta}} - \widehat{\underline{\theta}})$.

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. We also 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. The bootstrap confidence interval $\text{CI}^1_{0.95}$ is also usually conservative, with a true coverage rate lying mostly between 0.92 and 1. This is expected, since with our DGP $\theta_0 \notin \{\underline{\theta}_0, \overline{\theta}_0\}$, so that the asymptotic coverage is 1. In the online appendix, we obtain similar results for other DGP's and show that the bootstrap test for the binomial mixture model performs well in practice.

Finally, we provide some evidence regarding the computational cost of our method. Two cases should be distinguished. When $\widetilde{P} \in \mathcal{P}$, which can be tested simply as explained in Subsection D.1 of the online appendix, the maximum likelihood estimator is trivial to compute since $\widehat{P} = \widetilde{P} = (N_1/n, ..., N_K/n)'$. However, the bounds can be costly to obtain in this case. Table 3 shows that computing the bounds based on Equations (3.3) and (3.4) is actually quick for small K, but becomes demanding for high K. On the other hand, it is almost immediate for any K when we can rely on Equation (3.5), as is the case with the Theil index. Conversely, when $\widetilde{P} \notin \mathcal{P}$, the bounds can be computed at almost no cost in view of (3.2), but the computation of $\widehat{P}$, based on Lemma 3.1, becomes the bottleneck in terms of CPU. As discussed above, this case prevails when $K \geq 10$ for typical sample sizes. The first row of Table 3 shows that the corresponding time increases with K, which makes sense because the dimension over which we optimize increases, but remains very manageable even with $K = 20$. Finally, computing bootstrap confidence intervals is, as expected, much more expensive because we have to go through these steps many times (200 in our simulations).

Table 3: Elapsed CPU time for the estimation of the bounds and confidence intervals (in 100th of seconds)

	K				
	3	6	9	12	20
Constrained ML, $n = 100$	15.8	19.9	22.7	37.3	49.9
Theil bounds, Chebyshev, $n = \infty$	0.13	0.08	0.08	0.08	0.13
Theil bounds, regular, $n = \infty$	12.6	10.5	37.6	78.5	6,940
Duncan bounds, regular, $n = \infty$	9.2	25.0	62.2	223.6	6,055
CI, $n = 100$ (in seconds)	67.2	120.9	126.7	222.5	376.3

Note: the times reported in the table are average elapsed CPU times over 100 simulations. The DGP is the same as in Table 2. The first row corresponds to the time required to obtain $\hat{P}$ when $\tilde{m} \notin \mathcal{M}$. In rows 2 to 4, we let $n = \infty$ in the sense that $\tilde{P} = P_0$. Row 2 displays the CPU time needed to compute the bounds of the Theil by the Chebyshev method, following Equation (3.5). Rows 3 and 4 display the CPU time needed to compute the Theil and Duncan bounds following Equations (3.3) and (3.4). The last row displays the CPU time needed to compute the confidence intervals of both the Theil and Duncan indices, with 200 bootstrap iterations.

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 requires to be able to measure workplace segregation. Early works focused on gender or race segregation across occupations or industries, see e.g. Fields and Wolff (1991). Groshen (1991) is the first contribution to use the information available at the scale of establishments. Carrington and 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 and Maskin (1996) and Kramarz et al. (1996) analyze, in the US and the French cases, how skill dispersion, measured by 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 attempt to correct the indices.[18] Carrington and Troske (1997) present new results on black/white segregation introducing their method to correct for small-unit bias. Hellerstein and 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. Åslund and Skans (2010) and Glitz (2014) also use Carrington and Troske's method to attempt to compute workplace segregation in Sweden and in Germany.

In this section, we aim at computing the Theil and Duncan indices 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 them? 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 estimates of the Theil and Duncan indices. As a matter of comparison, we also display the naive estimate and the ones proposed by Carrington and Troske (1997), Allen et al. (2015) and Rathelot (2012).

We rely on the 2007 *Déclarations Annuelles de Données Sociales* (DADS), the French matched employer-employee database, which is exhaustive on the private sector (1.77 million establishments). In what follows, we restrict the sample to the 1.04 million establishments that have between 2 and 25 employees. We define the minority group as individuals born abroad and with the nationality of a country outside Europe. 3.7% of workers are considered as minority workers in the total population. We distinguish two categories of jobs: the least-skilled category gather white-collar unskilled jobs (*employés*) and blue-collar jobs (*ouvriers*). The other occupations form the skilled category. 41% of jobs belong to the unskilled category. While 40.7% of majority workers work in unskilled jobs, this is the case for 57.4% minority workers. Regressing the net wages of each worker on his and the job's characteristics, we check the economic relevance of our categories. We find that, conditional on sex, age and the number of days in the year, workers in unskilled jobs earn 29% less than those in skilled jobs, minority workers earn 8% less than

[18] Kremer and 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.

majority workers and being a minority worker in an unskilled job is associated with an additional penalty of 1.6%.

Before presenting our results, we first check that the binomial mixture model is not rejected in these data. For that purpose, we use the test that we consider in Subsection C.1 of the online appendix. For $K = 2...8$, $\widetilde{P} = \widehat{P}$, so the test is automatically accepted. For $K \geq 9$, $\widetilde{P} \neq \widehat{P}$, but this may be expected even if $P_0 \in \mathcal{P}$ given the results of our Monte Carlo simulations (see Figure 2 above). Performing the bootstrap test detailed above for $K \geq 8$, we do not reject the binomial mixture model at the 10% level for any value of K (see Table 4). We see this as evidence that the binomial mixture model is reasonable here.

Table 4: Test of the binomial mixture model.

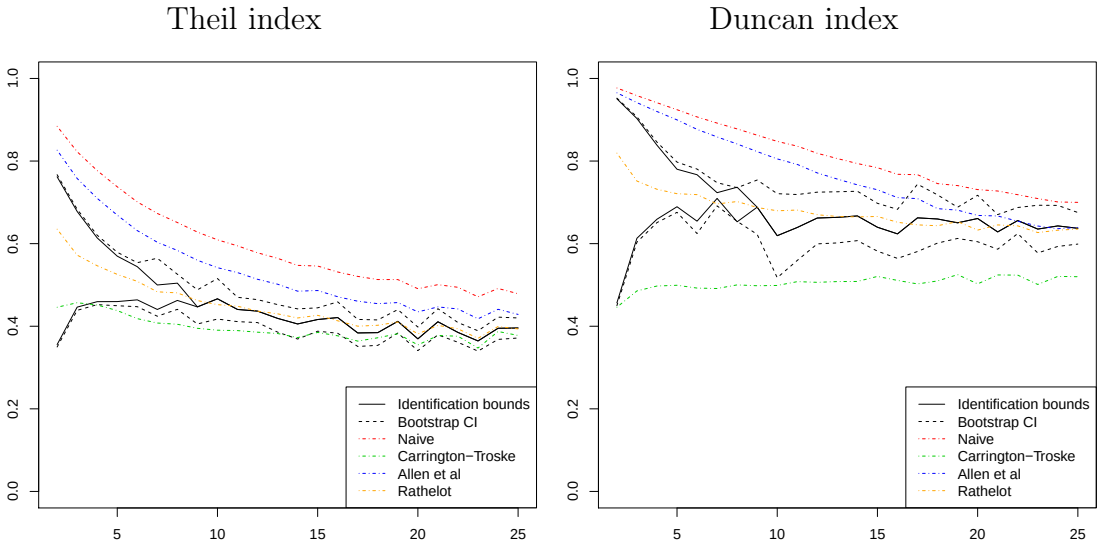
Unit size K	p-value of the bootstrap test	Unit size K	p-value of the bootstrap test
≤ 8	1	17	0.54
9	0.80	18	0.34
10	0.56	19	0.11
11	0.98	20	0.61
12	0.80	21	0.19
13	0.72	22	0.07
14	0.77	23	0.73
15	0.99	24	0.17
16	0.49	25	0.37

Note: for $K \leq 8$, $\widetilde{P} = \widehat{P}$, so that $LR_n = 0$ and p-value$= 1$.

Figure 3 displays the estimates of workplace segregation for different firm sizes, using the Theil and Duncan indices across French establishments. 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 $K \geq 9$, as expected since for these values, $\widetilde{P} \neq \widehat{P}$. Both for the Theil and the Duncan, the naive estimator is well above the upper bound of the 95% confidence interval. Carrington and Troske's correction works quite well for the Theil index, remaining inside the 95% confidence interval or close to its lower bound. However, in line with Figure 1, it strongly underestimates the Duncan index, the difference with our point estimate lying between 0.10 and 0.15 for $K \geq 9$. We observe a reversed pattern for the Allen et al. estimator. Their corrected Theil remains outside

the 95% confidence intervals for all unit sizes, while their corrected Duncan is close to our point estimate and mostly within the confidence interval for $K \geq 14$. The method proposed by Rathelot (2012) seems to perform well here for both indices, suggesting that the mixture of two beta distributions is a reasonable approximation for the distribution of p.

Figure 3: Theil and Duncan indices, by firm size.



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, much less so in the second case. The negative correlation between the index and the unit size is not surprising for the naive and the Allen et al. estimates, as the magnitude of their bias decrease with K (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. For instance, small firms may rely more heavily on social networks in their hiring process, resulting in a higher segregation between firms (people from the minority tending to hire other people from the same minority, and conversely).[19]

To test for this correlation, we consider the null hypothesis that $K \mapsto \theta_0(K)$ is

[19]Pistaferri (1999) shows that, in Italy, smaller firms tend to use more often informal hiring channels. In a similar vein, Giuliano et al. (2009) show, for the US, that manager's race affects the racial composition of new hires.

constant over $\mathcal{K}$, where $\theta_0(K)$ is the true parameter corresponding to firms of size K and $\mathcal{K}$ is a subset of firm sizes. Because of partial identification, developing such a test is not trivial, see Subsection C.2 of the online appendix for details. We consider three subsets $\mathcal{K}$ here: the whole range $\{2, ..., 25\}$, $\{2, ..., 9\}$, which corresponds to the definition of very small firms in France, and $\{10, ..., 25\}$. The results are displayed in Table 5. For both the Duncan and Theil indice, we do not reject the null hypothesis that $K \mapsto \theta_0(K)$ is constant on very small firms. We also accept at the 10% level the hypothesis of a constant Duncan index on $\{10, ..., 25\}$. We perform the same tests with the alternative methods (naive, Carrington and Troske's correction and Allen et al.'s correction), using the asymptotic normality of the corresponding estimators and estimating the asymptotic variance with bootstrap. For the three methods, three possible subsets $\mathcal{K}$ and two indices, we always reject the null hypothesis at the 1% level. Contrary to ours, these approaches do not satisfy the size invariance axiom mentioned above, which might cause the apparent dependence of segregation in K.

Table 5: Equality tests of segregation indices across unit sizes.

$\mathcal{K}$	Theil index	Duncan index
$K \leq 25$	$< 2.10^{-3}$	$< 2.10^{-3}$
$K \leq 9$	0.45	0.25
$10 \leq K \leq 25$	$< 2.10^{-3}$	0.11

Note: we use the subsampling test detailed in Subsection C.2 of the online appendix, with 500 subsamples.

Finally, we compute the bounds on the segregation indices for the whole set of firms. Results are displayed in Table 6. When considering the worker level and thus using the weighted index, we estimate the bounds to be [0.428, 0.514] on the Theil index and [0.634, 0.740] on the Duncan index. The uncertainty is thus quite large, a result mostly driven by the lack of information on very small firms, which represent a large proportion of our sample (83% of the firms are of size less than 9). When the index is unweighted, even more importance is given to the small firms and the identication set is wider. Because of the very large number of observations, the confidence intervals are not much wider than the identification sets in this case. For both the Theil and the Duncan, the naive and Allen et al.

estimates are above the upper bound, while the Carrington-Troske estimates are below the lower bound. The index corrected by the Rathelot method is just below the lower bound for the Theil but within the bounds for the Duncan.

Table 6: Comparison between the sharp bounds, the naive approach and previous corrections on all firms.

Method	Theil index		Duncan index	
	Estimate	$CI_{0.95}$	Estimate	$CI_{0.95}$
Sharp bounds				
Weighted	[0.428, 0.514]	[0.423, 0.521]	[0.634, 0.740]	[0.620, 0.746]
Unweighted	[0.423, 0.604]	[0.419, 0.609]	[0.596, 0.819]	[0.584, 0.823]
Alternative methods				
Naive	0.749	[0.747, 0.750]	0.915	[0.914, 0.915]
Carrington-Troske	0.421	[0.418, 0.424]	0.502	[0.500, 0.504]
Allen et al.	0.685	[0.683, 0.686]	0.888	[0.888, 0.889]
Rathelot	0.425	[0.421, 0.428]	0.659	[0.654, 0.661]
Conditional on job skill level				
Unskilled	[0.425, 0.514]	[0.422, 0.524]	[0.628, 0.730]	[0.614, 0.732]
Skilled	[0.423, 0.543]	[0.415, 0.551]	[0.620, 0.774]	[0.602, 0.782]
Average ($\theta_{0.}$)	[0.424, 0.532]	[0.418, 0.541]	[0.623, 0.757]	[0.607, 0.763]

Note: the conditional indices correspond to the θ_{0w} defined in Subsection B.2, while $\theta_{0.}$ is a weighted average of those two.

We have shown that minority workers are disproportionately represented in unskilled positions. Because the proportion of unskilled positions varies across firms, we can imagine that this simple correlation would increase the segregation of the minority across firms. The last rows of Table 6 show that this is not the case. First, segregation seems to remain of the same magnitude once we restrict our sample to either skilled or unskilled positions within the firm. Second, when we consider the average conditional index, which is the weighted average of the indices on the two types of positions, we also find that segregation remains at a very similar level. While the descriptive results mentioned above make us confident that the job category we have built are economically sensible, our results show that workplace segregation of minority does not merely reflect the higher share of

unskilled jobs among minority workers and the uneven distribution of unskilled positions across establishments. Rather, the same level of segregation seems to exist for both types of jobs.

Our result is in line with Hellerstein and Neumark (2008). Using a correction à la Carrington and Troske, they find that ethnic workplace segregation in the U.S. is not accounted for by differences in education across ethnic groups. However, they find that taking the language spoken into account explains an important part of segregation. In the case of France, we conjecture that language will not play as an important role as in the U.S. but we lack the appropriate data to test this hypothesis. Similarly, Åslund and Skans (2010) show that controlling for human capital does not affect much the segregation index in the case of ethnic workplace segregation in Sweden, using again an extended version of Carrington and Troske's correction.

6 Conclusion

In this paper, we investigate what can be learned on segregation indices when only an imperfect measure of p, distributed according to a binomial variable $B(K, p)$, is available. We show that in general this leads to partial identification of the segregation index. We then develop inference on the bounds. We have not considered here segregation indices that do not take the form imposed by Assumption 2.1, such as the Gini index. Optimizing over distributions with finite support, as done here, leads to bounds that are in general strictly included in the sharp identified set. To obtain valid confidence intervals, a solution would be to choose a number of points in the support large compared to the sample size, so that this problem becomes negligible compared to the sample variability.

Given their initial purposes, we believe that segregation indices should be functions of F_p. This does mean, however, that when studying segregation, focusing on p rather than on X (or X/K) is always preferable.[20] When concern is for the consequences of segregation, the distribution of interest might be the one of the realized shares. In the school or the residential context, the question is often about how some groups affect others' decisions and outcomes. For instance, in the school context, an important issue is how low- and high-performing students

[20]We thank an anonymous referee for his detailed suggestions about this aspect.

affect each other in a classroom and whether more or less segregation is desirable from an aggregate point of view. Along this line, Bhattacharya (2009) investigates how the actual allocation has to be modified in order to maximize an aggregate measure of welfare. Similarly, Graham et al. (2010) aims to estimate the impact on the average outcome of a change in the allocation of individuals, increasing or decreasing actual segregation.

Even if one aims at understanding the causes of segregation, the distribution of X/K may matter, depending on the theoretical model we consider. Specifically, suppose that when doing their choice of firms (or neighborhood), individuals value the composition of the firm in terms of the minority (for such an analysis on urban segregation, see Kasy, 2015). If they observe the actual composition, then X/K would matter as well. If not because, e.g., all individuals choose simultaneously as in Kasy (2015), then F_p is more an object of interest.

A Proofs

A.1 Proof of Proposition 2.3

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

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

Hence, F_p dominates stochastically $F_{X/K}$ at the second order, and by monotonicity, $g(F_p, m_{01}) \leq \theta_N$. Moreover, this is true for any distribution $F_p \in \mathcal{D}_{m_0}$ since such distributions rationalize the one of X/K. Choosing a sequence $(F_{n,p})_{n\in\mathbb{N}}$ in $\mathcal{D}_{m_0}$ such that $\lim_{n\to\infty} g(F_{n,p}, m_{01}) = \overline{\theta}_0$, we thus get $\overline{\theta}_0 \leq \theta_N$. When the support of p is not reduced to $\{0,1\}$, X/K is not a deterministic function of p with probability equal to one. Hence, for any strictly concave function u, the event $E\left[u\left(X/K\right)|p\right] < u\left(E\left[X/K|p\right]\right)$ holds with a positive probability. As a result, $E\left[u\left(X/K\right)\right] < E[u(p)]$, and the result follows by strict monotonicity of $g(., m_{01})$ □

A.2 Proof of Lemma 3.1

By Theorems III.4.1 and III.5.1 of Krein and Nudel'man (1977):

$$\mathcal{M} = \left\{ \left(\int x dF, ..., \int x^K dF \right)', F \in \mathcal{D}^{L+1} \right\}.$$

In other words, for any $m \in \mathcal{M}$, there exists a distribution with only $L+1$ support points that rationalize this distribution. This implies that $P = (P_1, ..., P_K)' \in \mathcal{P}$ if and only if there exists $(x, y) \in \mathcal{S}_{L+1} \times \mathcal{T}_{L+1}$ such that

$$P_k = Q_k \left(\sum_{j=1}^{L+1} y_j x_j^1, ..., \sum_{j=1}^{L+1} y_j x_j^K \right)'.$$

Using the definition of Q, we obtain after some algebra

$$P_k = \binom{K}{k} \sum_{j=1}^{L+1} y_j x_j^k (1 - x_j)^{K-k}, \quad k \in \{1, ..., K\}.$$

The result follows.

A.3 Proof of Theorem 3.1

We first establish the asymptotic distribution of $\widehat{P}$, before turning to the bounds. The unconstrained maximum likelihood estimator $\widetilde{P}$ is simply the vector of sample proportions $(N_1/n, ..., N_K/n)$. Therefore, by the central limit theorem,

$$\sqrt{n}\left(\widetilde{P} - P_0\right) \xrightarrow{d} \mathcal{Z},$$

where $\mathcal{Z} \sim \mathcal{N}(0, \Sigma(P_0))$. Now, the constrained maximum likelihood estimator $\widehat{P}$ satisfies $\widehat{P} = \rho(\widetilde{P})$, where ρ is defined as in Lemma F.1. Therefore, by this lemma,

$$\sqrt{n}\left(\widehat{P} - P_0\right) = \pi_{\overline{C}_{P_0}}\left(\sqrt{n}\left(\widetilde{P} - P_0\right)\right) + o_P(1).$$

By continuity of the projection, we obtain

$$\sqrt{n}\left(\widehat{P} - P_0\right) \xrightarrow{d} \pi_{\overline{C}_{P_0}}(\mathcal{Z}).$$

As a result,

$$\sqrt{n}\left(\widehat{m} - m_0\right) \xrightarrow{d} Q^{-1}\pi_{\overline{C}_{P_0}}(\mathcal{Z}). \tag{A.1}$$

Hence $\widehat{m}$ is consistent. $\underline{\theta}$ and $\overline{\theta}$ are continuous by Lemma F.2. Consistency of the estimated bounds follows by the continuous mapping theorem. The asymptotic distribution of the bounds also follows from (A.1) and the extended delta method of Shapiro (1991) □

A.4 Proof of Theorem 3.2

The proof consists in five steps.

1. Asymptotic normality of $\widetilde{P}^*$. Our bootstrap consists of drawing a i.i.d. sample $(X_1^*, ..., X_n^*)$ with

$$(\Pr(X_i^* = 1), ..., \Pr(X_i^* = K))' = \widehat{P}_b.$$

Moreover, introducing the function $I(x) = (\mathbb{1}\{x = 1\}, ..., \mathbb{1}\{x = K\})'$, we have $\widetilde{P}^* = \frac{1}{n}\sum_{i=1}^n I(X_i^*)$. Fix $\varepsilon > 0$. For n large enough, $\|I(X_i^*)\| \leq \varepsilon\sqrt{n}$. Therefore,

$$\frac{1}{n}\sum_{i=1}^n E\left[\|I(X_i^*)\|^2 \mathbb{1}\left\{\|I(X_i^*)\| > \varepsilon\sqrt{n}\right\}\right] \to 0.$$

Besides,

$$V\left(I(X_i^*)|\widehat{P}_b\right) = \Sigma(\widehat{P}_b) \xrightarrow{P} \Sigma(P_0).$$

Hence, by the Lindeberg-Feller central limit theorem (see, e.g., van der Vaart, 2000, Theorem 2.27), we have, conditional on $\widehat{P}_b$ and with probability approaching one,

$$\sqrt{n}\left(\widetilde{P}^* - \widehat{P}_b\right) \xrightarrow{d} \mathcal{N}\left(0,\, \Sigma(P_0)\right). \tag{A.2}$$

2. Asymptotic distribution of $\widehat{P}^*$. We now prove that

$$\sqrt{n}\left(\widehat{P}^* - \widehat{P}_b\right) \xrightarrow{d} \pi_{\overline{C}_{P_0}}(\mathcal{Z}), \tag{A.3}$$

where $\mathcal{Z} \sim \mathcal{N}\left(0, \Sigma(P_0)\right)$.

First, suppose that $P_0 \in \overset{\circ}{\mathcal{P}}$. Then with probability approaching one, $\widehat{P}_b \in \overset{\circ}{\mathcal{P}}$ and thus also $\widetilde{P}^* \in \mathcal{P}$. As a result, with probability approaching one, $\widehat{P}^* = \widetilde{P}^*$. Thus, (A.2) also holds when replacing $\widetilde{P}^*$ by $\widehat{P}^*$. (A.3) follows by remarking that $\overline{C}_{P_0} = \mathbb{R}^K$, so that $\pi_{\overline{C}_{P_0}}(\mathcal{Z}) = \mathcal{Z}$.

Next, suppose that $P_0 \in \partial\mathcal{P}$. Let $\mathcal{Z}_n^* = \sqrt{n}\left(\widetilde{P}^* - \widehat{P}_b\right)$. By the continuous mapping theorem, $\pi_{\overline{C}_{P_0}}(\mathcal{Z}_n^*) \xrightarrow{d} \pi_{\overline{C}_{P_0}}(\mathcal{Z})$. Therefore, it suffices to prove that

$$\sqrt{n}\left(\widehat{P}^* - \widehat{P}_b\right) - \pi_{\overline{C}_{P_0}}(\mathcal{Z}_n^*) \xrightarrow{P} 0. \tag{A.4}$$

For that purpose, remark that by Lemma F.1,

$$\begin{aligned}
\sqrt{n}\left(\widehat{P}^* - \widehat{P}_b\right) &= \sqrt{n}\left(\widehat{P}^* - P_0\right) + \sqrt{n}\left(P_0 - \widehat{P}_b\right) \\
&= \pi_{\overline{C}_{P_0}}\left(\sqrt{n}(\widetilde{P}^* - P_0)\right) + \sqrt{n}\left(P_0 - \widehat{P}_b\right) + o_P(1). \qquad \text{(A.5)}
\end{aligned}$$

By Assumption 3.3, the boundary $\partial\overline{C}_{P_0}$ of $\overline{C}_{P_0}$ is linear. Thus, it is the tangent space of $\mathcal{P}$ at P_0, and by definition,

$$\left\|\widehat{P}_b - \pi_{\partial\overline{C}_{P_0}}(\widehat{P}_b)\right\| = o_P\left(\left\|\widehat{P}_b - P_0\right\|\right).$$

Let $\pi_{\partial\overline{C}_{P_0}}$ denotes the linear projection onto the tangent space $\partial\overline{C}_{P_0}$ of $\overline{C}_{P_0}$ and $u_n = \pi_{\partial\overline{C}_{P_0}}(\sqrt{n}(P_0 - \widehat{P}_b))$. We get

$$\begin{aligned}
\left\|\sqrt{n}\left(P_0 - \widehat{P}_b\right) - u_n\right\| &= \sqrt{n}\left\|\widehat{P}_b - \pi_{\partial\overline{C}_{P_0}}(\widehat{P}_b)\right\| \\
&= \sqrt{n}\, o_P\left(\left\|\widehat{P}_b - \widetilde{P}\right\| + \left\|\widetilde{P} - P_0\right\|\right) \\
&= o_P\left(\sqrt{n}\left\|\widetilde{P} - P_0\right\|\right) = o_P(1), \qquad \text{(A.6)}
\end{aligned}$$

where the first equality stems from linearity of $\pi_{\partial\overline{C}_{P_0}}$, the second from the triangular inequality and the third from $\left\|\widehat{P}_b - \widetilde{P}\right\| = \min_{P\in\partial\mathcal{P}}\left\|P - \widetilde{P}\right\|$. Combining (A.5) and (A.6) yields

$$\sqrt{n}\left(\widehat{P}^* - \widehat{P}_b\right) = \pi_{\overline{C}_{P_0}}\left(\sqrt{n}(\widetilde{P}^* - P_0)\right) + u_n + o_P(1). \tag{A.7}$$

Now, remark that

$$\pi_{\overline{C}_{P_0}}(h) = h\mathbb{1}\{h \in \overline{C}_{P_0}\} + \pi_{\partial\overline{C}_{P_0}}(h)\mathbb{1}\{h \notin \overline{C}_{P_0}\}.$$

Besides, for all $h_1 \in \mathbb{R}^K$ and $h_2 \in \partial\overline{C}_{P_0}$, $h_1 + h_2 \in \overline{C}_{P_0}$ if and only if $h_1 \in \overline{C}_{P_0}$. As a result, for all $h_1 \in \mathbb{R}^K$ and $h_2 \in \partial\overline{C}_{P_0}$,

$$\pi_{\overline{C}_{P_0}}(h_1) + h_2 = \pi_{\overline{C}_{P_0}}(h_1 + h_2). \tag{A.8}$$

Hence,

$$\begin{aligned}\sqrt{n}\left(\widehat{P}^* - \widehat{P}_b\right) &= \pi_{\overline{C}_{P_0}}\left(\sqrt{n}(\widetilde{P}^* - P_0) + u_n\right) + o_P(1) \\ &= \pi_{\overline{C}_{P_0}}\left(\mathcal{Z}_n^*\right) + o_P(1),\end{aligned} \tag{A.9}$$

where the first equality follows by (A.7), (A.8) and the fact that $u_n \in \partial\overline{C}_{P_0}$, and the second by (A.6) and the fact that projections are continuous. (A.4), and therefore (A.3) follows.

3. Asymptotic distribution of $(\overline{T}^*, \underline{T}^*)$. We have $\widehat{m}^* = Q^{-1}\widehat{P}^*$. Moreover, $(\underline{\theta}, \overline{\theta})$ is differentiable at m_0. Applying the delta method for the bootstrap (see, e.g. van der Vaart, 2000, Theorem 23.9) then yields

$$\begin{pmatrix} \overline{T}^* \\ \underline{T}^* \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \overline{\theta}'(m_0)Q^{-1}\pi_{\overline{C}_{P_0}}(\mathcal{Z}) \\ \underline{\theta}'(m_0)Q^{-1}\pi_{\overline{C}_{P_0}}(\mathcal{Z}) \end{pmatrix}. \tag{A.10}$$

4. Asymptotic validity of the confidence interval when $m_0 \in \overset{\circ}{\mathcal{M}}$. When $m_0 \in \overset{\circ}{\mathcal{M}}$, $I_n \xrightarrow{P} 0$ and it suffices to show that

$$\inf_{\theta_0 \in [\underline{\theta}_0, \overline{\theta}_0]} \lim_{n\to\infty} \Pr\left(\theta_0 \in \mathrm{CI}_{1-\alpha}^{\text{interior}}\right) = 1 - \alpha. \tag{A.11}$$

Suppose first that $\theta_0 = \underline{\theta}_0$. Then

$$\begin{aligned}&\Pr\left(\theta_0 \in \mathrm{CI}_{1-\alpha}^{\text{interior}}\right) \\ &= \Pr\left(\underline{T} \le c_{1-\alpha}(\underline{T}^*),\ \overline{T} + \sqrt{n}(\overline{\theta}_0 - \underline{\theta}_0) \ge c_\alpha(\overline{T}^*)\right) \\ &= \Pr\left(\underline{T} \le c_{1-\alpha}(\underline{T}^*)\right) - \Pr\left(\underline{T} \le c_{1-\alpha}(\underline{T}^*),\ \overline{T} + \sqrt{n}(\overline{\theta}_0 - \underline{\theta}_0) < c_\alpha(\overline{T}^*)\right).\end{aligned} \tag{A.12}$$

Let P_1 and P_2 denote the two probability terms in (A.12). $\underline{\theta}$ is differentiable at m_0, with a nonzero gradient by Assumption 3.3. Besides, when $m_0 \in \overset{\circ}{\mathcal{M}}$, $\pi_{\overline{C}_{P_0}}(\mathcal{Z}) = \mathcal{Z}$. Thus, by Theorem 3.1, the asymptotic distribution of $\underline{T}$ is normal with strictly

positive variance. This distribution is therefore continuous at $c_{1-\alpha}(\underline{T})$. By Part 3 of the proof and Theorem 1.2.1 of Politis et al. (1999) (see also their remark 1.2.1), $P_1 \to 1-\alpha$ with probability one.

Besides, with probability one,

$$P_2 \leq \Pr\left(\overline{T} + \sqrt{n}(\overline{\theta}_0 - \underline{\theta}_0) < c_\alpha(\overline{T}^*)\right) \to 0,$$

since $c_\alpha(\overline{T}^*) = O_P(1)$ and $\sqrt{n}(\overline{\theta}_0 - \underline{\theta}_0) \to \infty$. As a result, with probability one,

$$\Pr\left(\theta_0 \in \mathrm{CI}_{1-\alpha}^{\mathrm{interior}}\right) \to 1-\alpha.$$

The same holds when $\theta_0 = \overline{\theta}_0$. Finally, if $\theta_0 \in (\underline{\theta}_0, \overline{\theta}_0)$,

$$\Pr\left(\theta_0 \in \mathrm{CI}_{1-\alpha}^{\mathrm{interior}}\right) = \Pr\left(\underline{T} + \sqrt{n}(\underline{\theta}_0 - \theta_0) \leq c_{1-\alpha}(\underline{T}^*),\ \overline{T} + \sqrt{n}(\overline{\theta}_0 - \theta_0) \geq c_\alpha(\overline{T}^*)\right).$$

Because $\underline{T} + \sqrt{n}(\underline{\theta}_0 - \theta_0) \to -\infty$ and $\overline{T} + \sqrt{n}(\overline{\theta}_0 - \theta_0) \to +\infty$, the probability on the right-hand side tends to one. Hence, (A.11) holds.

5. Asymptotic validity of the confidence interval when $m_0 \in \partial\mathcal{M}$. In this case, $\underline{\theta}_0 = \overline{\theta}_0 = \theta_0$. Thus, we have, by Theorem 3.1, $\sqrt{n}(\widehat{\overline{\theta}} - \widehat{\underline{\theta}}) = O_P(1)$. Because $k_n \to \infty$, $I_n \xrightarrow{P} 1$ and it suffices to show that with probability one,

$$\lim_{n\to\infty} \Pr\left(\theta_0 \in \mathrm{CI}_{1-\alpha}^{\mathrm{boundary}}\right) = 1-\alpha. \tag{A.13}$$

We have

$$\begin{aligned}\Pr\left(\theta_0 \in \mathrm{CI}_{1-\alpha}^{\mathrm{boundary}}\right) &= \Pr\left(\sqrt{n}\left|\widehat{\theta} - \frac{\underline{\theta}_0 + \overline{\theta}_0}{2}\right| \leq c_{1-\alpha}(T_s^*)\right) \\ &= \Pr\left(T_s \leq c_{1-\alpha}(T_s^*)\right),\end{aligned}$$

where in the first equality we have used the definition of $\mathrm{CI}_{1-\alpha}^{\mathrm{boundary}}$ and the fact that $\underline{\theta}_0 = \overline{\theta}_0 = \theta_0$. Remark that $T_s^* = |\underline{T}^* + \overline{T}^*|/2$. Thus, by Part 3 above, the continuous mapping theorem, Assumption 3.3 and Theorem 1.2.1 of Politis et al. (1999) once more,

$$\Pr\left(T_s \leq c_{1-\alpha}(T_s^*)\right) \to 1-\alpha$$

with probability one. The result follows □

References

Allen, R., S. Burgess, R. Davidson, and F. Windmeijer (2015): "More Reliable Inference for Segregation Indices," *Econometrics Journal*, 18, 40–66.

Andrews, D. K. (2000): "Inconsistency of the Bootstrap when a Parameter Is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–406.

Andrews, D. K. and S. Han (2009): "Invalidity of the Bootstrap and the m Out of n Bootstrap for Confidence Interval Endpoints Defined by Moment Inequalities," *Econometrics Journal*, 12, S172–S199.

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

Bayer, P. and R. McMillan (2012): "Tiebout sorting and neighborhood stratification," *Journal of Public Economics*, 96, 1129–1143.

Bhattacharya, D. (2009): "Inferring optimal peer assignment from experimental data," *Journal of the American Statistical Association*, 104, 486–500.

Brunello, G. and L. Rocco (2013): "The effect of immigration on the school performance of natives: Cross country evidence using PISA test scores," *Economics of Education Review*, 32, 234–246.

Card, D. and J. Rothstein (2007): "Racial segregation and the black-white test score gap," *Journal of Public Economics*, 91, 2158–2184.

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

——— (1997): "On Measuring Segregation in Samples with Small Units," *Journal of Business & Economic Statistics*, 15, 402–09.

——— (1998): "Interfirm Segregation and the Black/White Wage Gap," *Journal of Labor Economics*, 16, 231–60.

Chakravarty, S. and J. Silber (1994): "Employment Segregation Indices: An Axiomatic Characterization," in *Models and Measurement of Welfare and Inequality*, ed. by W. Eichhorn, Springer Berlin Heidelberg, 912–920.

Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2013):

"Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

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

Cutler, D. M. and E. L. Glaeser (1997): "Are Ghettos Good or Bad?" *The Quarterly Journal of Economics*, 112, 827–72.

Cutler, D. M., E. L. Glaeser, and J. L. Vigdor (1999): "The Rise and Decline of the American Ghetto," *Journal of Political Economy*, 107, 455–506.

——— (2008): "When are ghettos bad? Lessons from immigrant segregation in the United States," *Journal of Urban Economics*, 63, 759–774.

Dette, H. and K. Schorning (2013): "Complete classes of designs for nonlinear regression models and principal representations of moment spaces," *The Annals of Statistics*, 41, 1260–1267.

Dette, H. and W. J. Studden (1997): *The theory of canonical moments with applications in statistics, probability, and analysis*, vol. 338, John Wiley & Sons.

Echenique, F. and R. Fryer (2007): "A Measure of Segregation Based on Social Interactions," *Quarterly Journal of Economics*, 122, 441–85.

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

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013): "Long-Term Effects of Class Size," *The Quarterly Journal of Economics*, 128, 249–285.

Gentzkow, M. and J. M. Shapiro (2011): "Ideological Segregation Online and Offline," *The Quarterly Journal of Economics*, 126, 1799–1839.

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

Glitz, A. (2014): "Ethnic Segregation in Germany," *Labour Economics*, 29, 28–

40.

Graham, B., G. Imbens, and G. Ridder (2010): "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach," NBER Working Paper 16499.

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, 457–472.

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

Horn, R. A. and C. R. Johnson (1990): *Matrix analysis*, Cambridge University Press.

Hutchens, R. (2001): "Numerical measures of segregation: desirable properties and their implications," *Mathematical Social Sciences*, 42, 13–29.

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

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

Jahn, J., C. F. Schmid, and C. Schrag (1947): "The Measurement of Ecological Segregation," *American Sociological Review*, 12, 293–303.

James, D. R. and K. E. Taeuber (1985): "Measures of Segregation," *Sociological Methodology*, 15, 1–32.

Kasy, M. (2015): "Identification in a model of sorting with social externalities and the causes of urban segregation," *Journal of Urban Economics*, 85, 16–33.

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

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

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

Leckie, G. and H. Goldstein (2014): "A multilevel modelling approach to measuring changing patterns of ethnic composition and segregation among London secondary schools, 2001-2010," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

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

Massey, D. S. and N. A. Denton (1988): "The Dimensions of Residential Segregation," *Social Forces*, 67.

Pistaferri, L. (1999): "Informal Networks in the Italian Labor Market," *Giornale degli Economisti*, 58, 355–375.

Politis, D., J. Romano, and M. Wolf (1999): *Subsampling*, Springer.

Rathelot, R. (2012): "Measuring Segregation when Units are Small: a Parametric Approach," *Journal of Business & Economic Statistics*, 30, 546–553.

Romano, J. and A. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Shapiro, A. (1991): "Asymptotic analysis of stochastic programs," *Annals of Operations Research*, 30, 169–186.

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

Stoye, J. (2010): "Partial identification of spread parameters," *Quantitative Economics*, 1, 323–357.

van der Vaart, A. W. (2000): *Asymptotic Statistics*, Cambridge University Press.

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

Wood, G. R. (1999): "Binomial mixtures: Geometric Estimation of the Mixing

Distribution," *Annals of Statistics*, 27, 1706–1721.