Measuring Segregation on Small Units: A Partial Identification Analysis

Xavier D’Haultfœuille† and Roland Rathelot‡

First version: May 2011
This version: July 2013

Abstract

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

JEL Classification: C13, C14 and J71.

Keywords: segregation, small units, partial identification.

---

*We would like to thank Han Hong, an anonymous associate editor, two referees, Romain Aeberhardt and Mathias André and Clément de Chaisemartin for their comments. We also benefited from discussions with Christian Bontemps, Laurent Davezies, Guy Laroque, Adam Rosen and the participants of the ESEM in Oslo, 21st EC2 conference in Toulouse, Crest-LMI and INSEE-D3E seminars.

†CREST, xavier.dhaultfoeuille@ensae.fr.

‡CREST, roland.rathelot@ensae.fr.
1 Introduction

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

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

Several works propose solutions to deal with this issue. The most common way is to provide corrected versions of the indices, in an attempt to extract the signal from the noise. Winship (1977) has been the first to propose a corrected Duncan index. Carrington & Troske (1997) developed his idea and have also proposed an adjusted index, close to Winship’s. Allen et al. (2009) proposed a correction based on bootstrap. Finally, Rathelot (2012) develops an estimator which is consistent under a parametric condition on the underlying distribution. He also shows that his index and the one proposed by Allen et al. perform better than other solutions for many standard distributions on the underlying probability. However, none of these approaches is consistent for any distribution on this probability.

In this paper, we first reconsider the problem from an identification viewpoint. In line with the literature, we suppose that the observed number of people belonging to the minority group in a given unit follows, conditional on the unobserved probability, a binomial dis-

¹Measuring workplace segregation at the level of the firm has been recently featured in Carrington & Troske (1995), Carrington & Troske (1998a), Carrington & Troske (1998b), Hellerstein & Neumark (2008), Åslund & Skans (2010) or Glitz (2012). Likewise, there are recent attempts to measure segregation at the level of the schools or the classrooms; see Allen et al. (2009) or Söderström & Uusitalo (2010).
tribution. On the other hand, we remain completely agnostic on the distribution of the unobserved probability. The binomial assumption, which we show is testable, allows us to identify the first $K$ moments of the distribution of the unobserved probability, where $K$ denotes the size of the units. Because most of the existing segregation indices depend on the whole distribution of this probability, not only on its first moments, these indices are only partially identified in general. Bounds can be obtained by minimizing or maximizing these indices over distributions whose first moments match those identified in the data. This problem is 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 the popular 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. This result is related to a result of Chernozhukov et al. (2013) in the context of nonlinear panel data. In such models, bounds on marginal effects can be obtained by maximizing some functionals over the distribution of the fixed effect. Similarly to us, they show that one can actually restrict to discrete distributions with a low number of support points. When unit sizes vary, we also show how to combine the bounds on the subpopulations of same size to obtain bounds on the whole population of units. These bounds can be improved if the unobserved probability and the unit size are independent, a condition that is testable.

We also develop inference on the segregation index, using a two-step procedure. In the first step, we estimate the vector of moments by minimum distance under the constraint that it belongs to the set of moments of distributions on $[0, 1]$. The estimator takes a closed and very simple form whenever the constraint is slack. Otherwise, we use a characterization of the projection on this set to transform this problem into an optimization under only linear equality and inequality constraints. In the second step, the bounds are estimated by optimizing over finite-dimensional distributions whose first moments match the first-step estimator. Interestingly, the lower and upper bounds coincide when the constraint on the vector of moments is binding, 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. We also consider a confidence interval on the true identification interval that is uniformly conservative. Finally, we develop a bootstrap test of the binomial mixture
model, based on the distance between the unconstrained estimated vector of moment and
the constrained one.

Monte Carlo simulations indicate that our method works well for finite samples. They also
show that even for modest unit sizes \( K = 9 \), 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. Finally, contrary to what is
suggested by the naive or Allen et al. (2009) estimator, we cannot reject at standard levels
that there is no relationship between plant size and segregation.

The paper is organized as follows. Section two presents the binomial mixture model and
studies partial identification of parameters of interest in this model. Section three develops
inference on the bounds and a test of the binomial assumption. Monte Carlo simulations
are drawn in Section four, and the application to segregation in the workplace is displayed
in the fifth section. Section six concludes. The online appendix gathers additional Monte
Carlo simulations, supplementary details on the application and all proofs.

2 Identification

2.1 The setting

The population is assumed to be split into two groups, a group of interest and the rest of
the population, and to be distributed across units. Units may represent geographical areas,
classrooms, or, as in our application, firms. 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. This contrasts
with Allen et al. (2009) for example who let the size of the units tend to infinity while the
number of units is fixed.

The segregation index of the population of interest across units, \( \theta_0 \), is then a real functional
of \( F_p \). We make the following restriction on \( \theta_0 \). We let hereafter \( m_{01} = E(p) \).
Assumption 2.1 $\theta_0 = g(F_p, m_{01})$ where $g(\cdot, m_{01})$ is linear and $g$ is continuous when considering the weak convergence topology: if $F_n$ converges in distribution to $F$ and $m_n \to m_{01}$, $g(F_n, m_n) \to g(F, m_{01})$.

This assumption holds in particular for the popular Duncan (or dissimilarity) and Theil indices, which satisfy respectively

\[
D = \frac{1}{2} E \left[ \frac{p}{E(p)} - \frac{1-p}{E(1-p)} \right] = \int \frac{|u - m_{01}| dF_p(u)}{2m_{01}(1 - m_{01})},
\]

\[
T = 1 - \frac{E(p \ln(p))}{E(p) \ln(E(p))} = \int \frac{[m_{01} \ln(m_{01}) - u \ln(u)] dF_p(u)}{m_{01} \ln(m_{01})}.
\]

The linear restriction also applies to the coworker index, which satisfies $C = \int (u - m_{01})^2 dF_p(u)/(m_{01} - m_{01}^2)$. This index, based on the exposure/isolation indices developed by Bell (1954), has been recently used by Hellerstein & Neumark (2008) and Glitz (2012). On the other hand, our setting does not include the Gini index, which is a concave but not linear functional of $F_p$ ($G = (1 - m_{01} - \int F_p^2(u) du)/m_{01}$).

### 2.2 Identification with a fixed unit size

The main problem in the inference on $\theta_0$ is that $p$ is not observed directly. We only observe the size of the unit, $K$, and the number $X$ of individuals in that unit that belong to the group of interest. To ease the exposition, we first suppose that the size of units $K$ is constant, the case of random size being considered afterwards. Equivalently, our analysis can be seen conditional on the random variable $K$. 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 & Troske (1997) or Rathelot (2012). It is also asymptotically equivalent to the allocation mechanism considered by Allen et al. (2009) when the number of individuals and the number of units tend to infinity at the same rate. Finally, as we shall see below,
this assumption is testable.

Intuitively, 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 on $F_p$. Letting $m_{oi} = E(p^i)$, we have, after some algebra:

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

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

$$P_0 = Qm_0. \quad (2.1)$$

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 parameter $\theta_0$ depending only on $m_0$ is point identified. This is for instance the case of the coworker index $C$. Because $C = (m_{02} - m_{01}^2)/(m_{01} - m_{01}^2)$, it is point identified as soon as $K \geq 2$.

The second implication of (2.1) 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, the sharp lower and upper bounds $\underline{\theta}_0$ and $\overline{\theta}_0$ of the identified set of $\theta_0$ satisfy

$$\underline{\theta}_0 = \inf_{F \in \mathcal{P}_{m_0}} g(F, m_{01}), \quad \overline{\theta}_0 = \sup_{F \in \mathcal{P}_{m_0}} g(F, m_{01}), \quad (2.2)$$

where $\mathcal{P}_{m_0}$ is the set of cumulative distribution functions on $[0, 1]$ with first $K$ moments equal to $m_0$. The identified set of $\theta_0$ is the whole interval $[\underline{\theta}_0, \overline{\theta}_0].$\textsuperscript{5}

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

$$\underline{\theta}_{0, \ell} = \inf_{F \in \mathcal{P}_{m_0}^\ell} g(F, m_{01}), \quad \overline{\theta}_{0, \ell} = \sup_{F \in \mathcal{P}_{m_0}^\ell} g(F, m_{01}), \quad (2.3)$$

of individuals $n$ tends to infinity together with the number of units, such that $\pi_i n \rightarrow \rho_i$, then the number of individuals of group $g$ in unit $i$, $X_i^g$, follows a Poisson distribution with parameter $\rho_i^g$. 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.

\textsuperscript{5}To see this, let, for any $\epsilon > 0$, $E \in \mathcal{P}_{m_0}$ (resp. $F \in \mathcal{P}_{m_0}$) be such that $g(E, m_{01}) < \underline{\theta}_0 + \epsilon$ (resp. $g(F, m_{01}) > \overline{\theta}_0 - \epsilon$). Then, for any $t \in [0, 1]$, $tE + (1-t)F \in \mathcal{P}_{m_0}$, and by Assumption 2.1 $g(tE + (1-t)F, m_{01}) = tg(E, m_{01}) + (1-t)g(F, m_{01})$. This implies that the interval $[\underline{\theta}_0 + \epsilon; \overline{\theta}_0 - \epsilon]$ is included in the identified set of $\theta_0$.  

6
where $P_{m_0}^\ell$ denotes the subset of $P_{m_0}$ with at most $\ell$ points of support. Of course, because the optimization set is smaller, we only obtain inner bounds in general, i.e. $\bar{\theta}_{0,\ell} \geq \bar{\theta}_0$ and $\bar{\theta}_{0,\ell} \leq \bar{\theta}_0$. However, under the linearity condition on the index, Theorem 2.1 below shows that the inner bounds actually coincide with the sharp bounds for $\ell = K + 1$.

**Theorem 2.1** Suppose that Assumption 2.1 holds. Then $\theta_{0,K+1} = \bar{\theta}_0$ and $\bar{\theta}_{0,K+1} = \bar{\theta}_0$.

Note that $P_{m_0}^{K+1}$ can be seen as a subset of $[0,1]^{2K+1}$, as any $F \in P_{m_0}^{K+1}$ is defined by its support points and associated probabilities. As a result, $\bar{\theta}_{0,K+1}$ and $\bar{\theta}_{0,K+1}$ can be obtained as an optimization over a subset of $[0,1]^{2K+1}$. This result is related to Lemma 7 of Chernozhukov et al. (2013), which addresses the partial identification of marginal effects in the context of nonlinear panel data. The bounds on marginal effects can be obtained by maximizing a linear functional over distributions of the fixed effect. They prove, as here, that the computation of these bounds only require to optimize over discrete distributions with a low number of support points.

To prove the theorem, we first show, relying on the Krein-Milman theorem, a deep result in convex theory, that optimization can be conducted on $\text{Ext}(P_{m_0})$ only, the set of extreme distributions of $P_{m_0}$. We then use a result of Douglas (1964) to show that $\text{Ext}(P_{m_0}) \subset P_{m_0}^{K+1}$. Noteworthy, the proof for the lower bound also applies to concave functionals. Because $g(.,m_{01})$ is concave in the case of the Gini index, $\theta_{0,K+1} = \bar{\theta}_0$ for the Gini. However, the upper bound cannot be obtained similarly.

### 2.3 Identification with a random unit size

We now turn to the case where $K$ is random and takes values in $\{2, ..., K\}$. We exclude here units with one individual, for which bounds on segregation indices are typically trivial. Let $F^k_p$ denote the cdf of $p$ conditional on $K = k$. Because $F_p = \sum_{k=2}^{K} \text{Pr}(K = k)F^k_p$ and $g(.,m_{01})$ is linear,

$$\theta_0 = \sum_{k=2}^{K} \text{Pr}(K = k)\theta^k_0,$$

where $\theta^k_0 = g(F^k_p, m_{01})$ would be the segregation index for units of size $k$ if the expectation of $F^k_p$ was $m_{01}$ (but in general $m_{01} = E(p) \neq E(p|K = k)$). Therefore, without any joint restriction on $(F^k_p, F^\ell_p)$, the sharp bounds on $\theta_0$ satisfy

$$\bar{\theta}_0 = \sum_{k=2}^{K} \text{Pr}(K = k)\bar{\theta}^k_0, \quad \bar{\theta}_0 = \sum_{k=2}^{K} \text{Pr}(K = k)\bar{\theta}^k_0, \quad (2.4)$$

where the bounds $\theta^k_0$ and $\bar{\theta}^k_0$ can be computed using Theorem 2.1 since $K = k$ is fixed.
Such bounds are obtained when one is fully agnostic on the dependence between \( p \) and \( K \), which is a safe option in the cases in which unit size might be a potential determinant of segregation. However, if one is ready to impose independence between these two variables, we can use units of size \( K \) to identify the first \( K \) unconditional moments of \( p \). Actually, the vector of unconditional moments \( m_0 = (m_{01}, ..., m_{0K}) \) is overidentified by

\[
P_0 = Qm_0,
\]

where \( P_0 \) (resp. \( Q \)) stacks together vectors \( P_0^k = (\Pr(X = 1|K = k), ..., \Pr(X = k|K = k)) \) (resp. the matrices \( Q^k \) of typical element \( \binom{k}{j} \binom{j}{i} (-1)^{j-i} \)) for different \( k \). Besides, because \( F_p = F_p^k \) for all \( k \), Theorem 2.1 directly applies to \( \theta_0 \) and \( \bar{\theta}_0 \) rather than on \( \theta_{0k}^k \) and \( \bar{\theta}_{0k}^k \). Apart from the accuracy gains due to the overidentification of \( m_0 \), the bounds on \( \theta_0 \) are thus likely to be very close, since they exploit the knowledge of the first \( K \) moments, with \( K \) potentially large. In particular, \( \theta_0 \) is point identified when \( K = \infty \), because the knowledge of all moments of a distribution on \([0, 1]\) fully characterizes this distribution (see, e.g., Gut, 2005, Theorems 8.1 and 8.3), reducing \( P_{m_0} \) to a singleton. However, to avoid any incorrect inference, the independence assumption should not be used when the overidentification test based on Equation (2.5) is rejected.

Up to now, we have also supposed that all individuals in the unit were observed. A common situation, however, is that \( n_K < K \) individuals only are sampled from units. In this case, \( X \) denotes the number of individuals belonging to the group of interest in this subsample. As previously, \( X \) follows, conditional on \( p \) and \( n_K \), a binomial distribution \( B(n_K, p) \). Hence, the result above applies by replacing \( K \) by \( n_K \). The main difference with previously is that in this case, it is plausible to assume \( n_K \) to be independent of \( p \) conditional on \( K \), whereas the assumption that \( p \) is independent of \( K \) is stronger in general. Under this condition, the \( n_K \) first moments of \( p \) conditional on \( K \) are identified, where \( \pi_K \) denotes the maximum of the support of \( n_K \). Once again, Theorem 2.1 applies directly by replacing \( K \) by \( \pi_K \) and we can recover bounds on the segregation index for the whole population using (2.4), up to this change.

### 2.4 Testability

The binomial mixture model that we consider is testable. To see this, note that a vector of raw moments should satisfy some restrictions, such as variance positivity \( (m_{02} \geq m_{01}^2) \). These restrictions may not hold if the model is not binomial. For instance, supposing that \( K = 2 \), the vector \( P_0 = (0.6, 0.3)' \) would correspond to the vector of raw moments \( m_0 = (0.6, 0.3)' \) according to the binomial model. But \( 0.3 - 0.6^2 < 0 \), which violates the
restriction that a variance is positive. This implies that such a vector $P_0$ invalidates the binomial mixture model.

More generally, the issue of whether a given vector belongs to the set $M$ of first $K$ moments of a probability distribution on $[0, 1]$ is known as the truncated Hausdorff problem. Several necessary and sufficient conditions have been established for this problem. Proposition 2.2 below, which is proved for instance by Krein & Nudel’man (1977, Theorem III.2.4), provides a characterization that is rather simple to use. We let afterwards $L$ denote the integer part of $K/2$, and, for all $\mu = (\mu_1, ..., \mu_K)' \in [0, 1]^K$, $A_\mu$, $B_\mu$, $C_\mu$ denote the square matrices of size $L + 1$, $L + 1$ and $L$ respectively, with typical $(i, j)$ term equal to $\mu_{i+j-2}$, $\mu_{i+j-1}$ and $\mu_{i+j-1} - \mu_{i+j}$ respectively (with the convention that $\mu_0 = 1$).

Proposition 2.2 For any $\mu \in [0, 1]^K$, $\mu \in M$ if and only if:
- $A_\mu - B_\mu$ and $B_\mu$ are positive when $K$ is odd.
- $A_\mu$ and $C_\mu$ are positive when $K$ is even.

2.5 Comparison with other approaches

Broadly speaking, previous approaches have ignored the issue of partial identification. Rather, they have focused on the estimation of parameters that are identified, but are different from $\theta_0$ 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 $\theta_N = g(F_X, m_{01})$. However, if $g(., m_{01})$ is monotonous with respect to the second-order dominance, as is the case of most inequality indices (including the Duncan and the Theil), this parameter is always greater than $\bar{\theta}_0$.

Proposition 2.3 Suppose that $g(., m_{01})$ is decreasing with respect to the second-order dominance. Then $\theta_N \geq \bar{\theta}_0$. Moreover, the inequality is strict if $g(., m_{01})$ is strictly decreasing 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 & Troske (1997), based on earlier works by Winship (1977) and Cortese et al. (1978). Suppose here, and without loss of generality if $g$ is bounded, that $g$ ranges from 0 to 1, the corrected index $\theta_{CT}$ of Carrington & Troske (1997) is defined by

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

\textit{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 $g(F, m_{01}) > g(G, m_{01})$.}
where \( \theta^*_N = g(F_{X/K}, m_{01}) \) (and \( X^{ns} \sim B(E(p), K) \)) is the naive parameter that would be obtained if all units had the same probability, that is if there was no segregation. The index \( \theta_{CT} \) can be seen as an affine correction that is valid in the two polar cases where there is no segregation (because \( \theta_{CT} = \theta_N = \theta_0 = 0 \) in this case) or if segregation is maximal (because then \( \theta_{CT} = \theta_0 = 1 \)). However, in general \( \theta_{CT} \) does not lie inside the interval \([\theta_0, \theta_0]\), as Figure 1 below shows. This is because \( \theta_{CT} \) does not correspond in general to a segregation index \( g(F_p, m_{01}) \) with \( p \in P_{m_0} \).

Allen et al. (2009) 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 \( \theta_0 \) 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 \( X^*|X \sim B(K, X/K) \). 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 \( \theta_{ABW} \) decreases more quickly than the one of \( \theta_N \) as \( K \to \infty \) (see the online appendix).

Finally, Rathelot (2012) introduces another correction based on the parametric assumption that the distribution of \( p \) is 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 \([\theta_0, \theta_0]\). Importantly also, and contrary to the previous approaches, this corrected index does not converge in general to the true one as \( K \to \infty \), unless one also makes the number of components of the mixture tends to infinity with \( K \). Yet, this correction seems to work rather well in practice even in the case of misspecification.

Figure 1 presents a comparison, for the Theil and Duncan index, between the sharp bounds, the naive approach and the corrections proposed by Carrington & Troske (1997), Allen et al. (2009) and Rathelot (2012) when \( \Phi^{-1}(p) \sim N(\mu, \sigma^2) \). \( \mu \) and \( \sigma^2 \) were chosen so as to match the first two estimated moments of \( p \) in our application in Section 5. \( E(p) \approx 0.033 \)

\footnote{In their framework, \( \theta^*_N \) 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 4.}
and $E(p^2) \simeq 0.011$ lead to $\mu \simeq -3.26$ and $\sigma^2 \simeq 2.13$. The sharp bounds are obtained by solving (2.3), 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: $\Phi^{-1}(p) \sim N(-3.26, 2.13)$. With this DGP, the average size of the minority is $E(p) \simeq 0.033$, the Theil index is $T \simeq 0.543$ and the Duncan index is $D \simeq 0.779$.

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. (2009) or Carrington & 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 & Troske (1997) performs better with the Theil than with the Duncan index. With this particular data generating process, the parametric method of Rathelot (2012) lies within the bound for all $K \leq 10$. 

11
3 Estimation

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. Following the identification part, we estimate the identified set by estimating its sharp bounds. When unit sizes are fixed and equal to \(K\), Theorem 2.1 implies that these bounds are equal to \(\theta_{0,K+1}\) and \(\bar{\theta}_{0,K+1}\). We estimate them using (2.3), where the unknown vector \(m_0\) is replaced by a suitable estimator. The case of a random unit size is similar and considered at the end of the subsection.

To define this estimator, remark first that we can estimate \(P_0\) by \(\tilde{P} = (\tilde{P}_1, ..., \tilde{P}_K)\) with \(\tilde{P}_k = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i = k\}\), and use Equation (2.1) to obtain the simple estimator \(\tilde{m} = Q^{-1}\tilde{P}\).

\(\tilde{m}\) is simple to compute but does not necessarily belong to \(\mathcal{M}\). Our Monte Carlo simulations show that this issue occurs with probability close to one when \(K \geq 10\), even with sample sizes as large as 10,000 (see also Wood, 1999). In such cases, we cannot optimize on \(\mathcal{P}_{\tilde{m}}\) to estimate bounds since \(\mathcal{P}_{\tilde{m}}\) is empty. To overcome this issue, we consider afterwards the constrained minimum distance estimator \(\hat{m}\) defined by

\[
\hat{m} = \arg \min_{m \in \mathcal{M}} (\tilde{P} - Qm)'W(\tilde{P} - Qm),
\]

where \(W\) is a positive definite matrix. \(\hat{m}\) is well defined as the unique projection onto the closed convex set \(\mathcal{M}\). Besides, \(\mathcal{P}_{\hat{m}}\) is never empty by definition of \(\hat{m}\). To compute \(\hat{m}\) in practice, first check whether \(\tilde{m} \in \mathcal{M}\), using Proposition 2.2 above. If this is the case, \(\hat{m} = \tilde{m}\). If not, we could use Proposition 2.2 to obtain \(\hat{m}\), but this entails computational problems, due to a nonlinear optimization with nonlinear inequality constraints. We rather rely on the following proposition, which follows directly from Karlin & Shapley (1949) and Karlin & Schumaker (1967), given that \(\hat{m} \in \partial \mathcal{M}\), the boundary of \(\mathcal{M}\).

**Proposition 3.1** Suppose that \(\hat{m} \neq \tilde{m}\). Then \(\mathcal{P}_{\hat{m}}\) is reduced to a single distribution with at most \(L + 1\) points of support.

By Proposition 3.1, the unique distribution \(F_{\hat{m}}\) corresponding to \(\hat{m}\) can be described by an ordered vector \(\hat{x} = (\hat{x}_1, ..., \hat{x}_{L+1})' \in \mathcal{S}_{L+1} = \{(x_1, ..., x_{L+1}) : 0 \leq x_1 < ... < x_{L+1} \leq 1\}\) and a vector of corresponding probabilities \(\hat{y} = (\hat{y}_1, ..., \hat{y}_{L+1})' \in \mathcal{T}_{L+1} = \{(y_1, ..., y_{L+1}) \in \mathbb{R}^{L+1} : \sum y_i = 1, 0 \leq y_i \leq 1\}\).
[0, 1]^{L+1} : \sum_{k=1}^{L+1} y_k = 1\}. For all \( x \in \mathcal{S}_{L+1} \), let

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

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

\[
(\hat{x}, \hat{y}) = \arg \min_{(x, y) \in \mathcal{S}_{L+1} \times \mathcal{T}_{L+1}} (\tilde{P} - QB(x)y)'W(\tilde{P} - QB(x)y),
\]

and \(\hat{m} = B(\hat{x})\hat{y}\). As a result, the computation of \(\hat{m}\) involves an optimization over a low dimensional space \(2(L + 1)\) under linear equality and inequality constraints only. This optimization turns out to be fast in practice.

Proposition 3.1 has a second important consequence. When \(\tilde{m} \not\in \mathcal{M}\), \(\min_{F \in \mathcal{P}_m} g(F, \hat{m}) = \max_{F \in \mathcal{P}_m} g(F, \hat{m})\) because \(\mathcal{P}_m = \{F_{\tilde{m}}\}\). Estimators of the bounds on \(\theta_0\) are defined in this case by

\[
\hat{\theta} = \tilde{\theta} = g(F_{\tilde{m}}, \hat{m}).
\]

When \(\tilde{m} \in \mathcal{M}\), on the other hand, \(\mathcal{P}_m\) is not reduced to a single distribution, and some optimization is required. We estimate the bounds by the estimators of \(\tilde{\theta}_{0, K+1}\) and \(\theta_{0, K+1}\). Given a vector of moments \(m = (m_1, \ldots, m_K)\), any \(F \in \mathcal{P}_m^{K+1}\) is defined by its support points \(x = (x_1, \ldots, x_{K+1}) \in \mathcal{S}_{K+1}\) and the associated probabilities \(y = (y_1, \ldots, y_{K+1})' \in \mathcal{T}_{K+1}\). Moreover, the moment constraints write \(A(x)y = (1, m')'\), where \(A(x)\) is the Vandermonde matrix

\[
A(x) = \begin{pmatrix} 1 & \cdots & 1 \\
x_1 & \cdots & x_{K+1} \\
\vdots \\
x_1^K & \cdots & x_{K+1}^K
\end{pmatrix}.
\]

Such a Vandermonde matrix is nonsingular for any \(x \in \mathcal{S}_{K+1}\) (see, e.g., Horn & Johnson, 1990). Thus, the vector of probabilities \(y\) satisfies \(y = A(x)^{-1}(1, m')'\), and constraints are equivalent to \(A(x)^{-1}(1, m')' \geq 0\), where inequalities are understood componentwise.

Because \(F \in \mathcal{P}_m^{K+1}\) 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

---

8When the distribution that rationalizes \(\hat{m}\) has less than \(L + 1\) support points, Program (3.1) does not admit a unique solution because we can set some components of \(y\) to zero and move freely the corresponding components of \(x\). In this case any solution can be chosen, since they all lead to the same \(\hat{m}\).
when the vector of moments is \( m_0 = m \) satisfy

\[
\bar{\theta}(m) = \max_{x \in S_{K+1}: A(x)^{-1}(1,m') \geq 0} q(x,m), \tag{3.2}
\]

\[
\underline{\theta}(m) = \min_{x \in S_{K+1}: A(x)^{-1}(1,m') \geq 0} q(x,m). \tag{3.3}
\]

Our estimators of \( \bar{\theta} \) and \( \underline{\theta} \) are respectively \( \hat{\theta} = \bar{\theta}(\hat{m}) \) and \( \hat{\underline{\theta}} = \underline{\theta}(\hat{m}) \).

When \( K \) is random and independent of \( p \), we can proceed very similarly. We first estimate \( m_0 = (m_{01}, ..., m_{0K}) \) by minimum distance (see, e.g., Wooldridge, 2002), using (2.5), the empirical counterpart of \( P_0 \) and the constraint \( m_0 \in M \). The bounds can then be estimated as previously, with \( K \) simply replaced by \( K' \).

Finally, when \( K \) is random and may depend on \( p \), we estimate the bounds by taking the empirical counterpart of (2.4). We apply the same ideas as above. First, we estimate \( m_{01} = E(X/K) \) by the average \( \hat{m}_1 = \frac{1}{n} \sum_{i=1}^n \frac{X_i}{K_i} \). Second, considering for any \( k \)

\[
\hat{P}^k = \left( \frac{\sum_{i : K_i = k} \mathbb{1}\{X_i = 1\}}{\sum_{i = 1}^n \mathbb{1}\{K_i = k\}}, ..., \frac{\sum_{i : K_i = k} \mathbb{1}\{X_i = k\}}{\sum_{i = 1}^n \mathbb{1}\{K_i = k\}} \right),
\]

we let \( \hat{m}^k = (Q^k)^{-1} \hat{P}^k \) denote the initial estimator of the first \( k \) moments of \( F_p^k \). \( \hat{m}^k \) is then the projection of \( \hat{m}^k \) onto \( M \). Third, for any \( F_p^k \in P_{m_k}^{k+1} \) with support \( x^k \), we may rewrite \( g(F_p^k, m_{01}) \) as a function of \( x^k, m^k \) and \( m_{01} \) only. We denote this function \( q(x^k, m^k, m_{01}) \), and let

\[
\bar{\theta}^k(m^k, m_{01}) = \max_{x^k \in S_{K+1}: A(x^k)^{-1}(1,m') \geq 0} q(x^k, m^k, m_{01}),
\]

\[
\underline{\theta}^k(m^k, m_{01}) = \min_{x^k \in S_{K+1}: A(x^k)^{-1}(1,m') \geq 0} q(x^k, m^k, m_{01}).
\]

If \( \hat{\Pr}(K = k) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{K_i = k\} \), we then estimate the bounds by

\[
\hat{\theta} = \sum_{k=2}^K \hat{\Pr}(K = k) \bar{\theta}^k(\hat{m}^k, \hat{m}_1), \quad \hat{\underline{\theta}} = \sum_{k=2}^K \hat{\Pr}(K = k) \underline{\theta}^k(\hat{m}^k, \hat{m}_1). \tag{3.4}
\]

### 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 \( m_0 \in \partial \tilde{M} \) and \( m_0 \in \partial \tilde{M} \), since the corresponding asymptotic distributions differ. Our asymptotic result is based on Assumption 3.1 below. Hereafter, we let \( C(x, m) = A(x)^{-1}(1,m') \), \( C(x, m) = (C_1(x, m), ..., C_{K+1}(x, m))^t \) and \( \nabla C(x, m) = \partial C / \partial x(x, m) \). Finally, \( \nabla C(x, m) \) is the submatrix of \( \nabla C(x, m) \), picking only the lines \( i \) for which \( C_i(x, m) = 0 \). In other words, \( \nabla C(x, m) \) is the gradient of the constraints that are binding at \( x \).
Assumption 3.1 q is locally Lipschitz and Clarke regular. For all $m \in \mathcal{M}$, $\nabla C(\hat{x}, m)$ is full rank at every $\hat{x}$ such that $\theta(m) = q(\hat{x}, m)$ or $\bar{\theta}(m) = q(\hat{x}, m)$.

The regularity condition on $q$ in Assumption 3.1 is very mild. It is satisfied for the Theil index but also for the Duncan index, even though $q(x, m) = (1, m)A(x)^{-1}x - m_1)/(2m_1(1 - m_1))$ is not differentiable at $(x_0, m_0)$ whenever $x_{0k} = m_{01}$ for some $k \in \{1, \ldots, K\}$. $q$ is still locally Lipschitz but also Clarke regular. As for the full rank condition, it is automatically satisfied if Programs (3.2) and (3.3) admit unique solutions, as assumed in Assumption 3.3 below, and the corresponding probabilities are strictly positive. In such cases, there is no binding constraints since all probabilities associated to $x$ are strictly positive. Otherwise, it depends on the true $m_0$. Considering boundary cases with $K = 2$ for instance, one can show that it fails to hold when $m_0 = (m_{01}, m_{01}^2)$, or equivalently when $p$ is constant, while it is satisfied when $m_0 = (m_{01}, m_{01})$, corresponding to $p \sim$Bernoulli$(m_{01})$.

We establish in the proof of Theorem 3.2 that under Assumption 3.1, $\bar{\theta}$ and $\theta$ admit directional derivatives. We let $\bar{\theta}'(m_0, h), \theta'(m_0, h)$ denote these directional derivatives at $m_0$ in the direction $h$, for any $h \in \mathbb{R}^K$. Finally, let $\Sigma = Q^{-1} [\text{diag}(P_0) - P_0P_0']Q^{-1}$, $\text{diag}(P_0)$ being the diagonal matrix with diagonal vector equal to $P_0$.

Theorem 3.2 Suppose that Assumption 2.1 holds. Then $(\hat{\theta}, \bar{\theta}) \overset{d}{\rightarrow} (\theta, \bar{\theta})$. If Assumption 3.1 also holds, then

$$
\sqrt{N} \left( \frac{\hat{\theta} - \theta_0}{\sigma} \right) \overset{d}{\rightarrow} \left( \frac{\theta'(m_0, \pi_{C_{m_0}}(Z))}{\theta'(m_0, \pi_{C_{m_0}}(Z))} \right),
$$

where $Z \sim \mathcal{N}(0, \Sigma)$ and $\pi_{C_{m_0}}$ is the projection onto the closure of the convex cone $C_{m_0} = \{\lambda(m - m_0), \ m \in \mathcal{M}, \lambda > 0\}$.

The estimated bounds are thus consistent under the minimal requirement of Assumption 2.1, while the asymptotic distribution also hinges on Assumption 3.1. Importantly, these results apply whether or not $m_0$ lies in the interior of $\mathcal{M}$. Note that $\theta$ and $\bar{\theta}$ can be proved to be differentiable under stronger conditions on $q$, such as Assumption 3.2 below. In such cases, and if $m_0 \in \partial \mathcal{M}$, the estimated bounds are asymptotically normal, because $\pi_{C_{m_0}}(Z) = Z$ and $\theta'(m_0, \cdot)$ and $\bar{\theta}'(m_0, \cdot)$ are linear.

Because the estimated bounds are not asymptotically normal in general, and in particular when $m_0 \in \partial \mathcal{M}$, the confidence interval proposed by Imbens & Manski (2004) for partially

\footnote{A function $f : \mathbb{R}^m \rightarrow \mathbb{R}^p$ is Clarke regular if for all $x, h$ in $\mathbb{R}^m$, $f$ admits a directional derivative $f'(x, h)$ at $x$ in the direction $h$ and $f'(x, h) = \limsup_{t \rightarrow 0^+} \frac{f(y + th) - f(y)}{t}$.}
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 \( \hat{m} \) onto \( \partial \mathcal{M} \) whenever \( \hat{m} \in \hat{\mathcal{M}} \) but is close to the boundary. Let \( d_n = \sqrt{n}(\hat{\theta} - \bar{\theta})/k_n \) and \( I_n = 1 \{d_n \leq 1\} \), with \( k_n \to \infty \), \( \sqrt{n}/k_n \to \infty \). Observe that when \( m_0 \in \partial \mathcal{M} \), \( \hat{\theta} = \bar{\theta} \) so that \( d_n \xrightarrow{P} 0 \) and \( I_n \xrightarrow{P} 1 \). When \( m_0 \in \mathcal{M} \) on the other hand, \( \hat{\theta} \neq \bar{\theta} \) 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

\[
\hat{m}_b = \pi_{\partial \mathcal{M}}(\hat{m})I_n + \hat{m}(1 - I_n),
\]

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

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 \( m_0 \in \mathcal{M} \). For any statistic \( T \), let \( T^* \) denote the corresponding bootstrap statistic. For example, if \( T = \sqrt{n}(\hat{\theta} - \theta) \), we let \( T^* = \sqrt{n}(\hat{\theta}^* - \theta) \), where \( \hat{\theta}^* \) is the bootstrap estimator of \( \theta \). We let \( c_{\alpha}(T^*) \) denote the \( \alpha \)-th quantile of the distribution of \( T^* \) conditional on \( \hat{m}_b \). We first define a confidence interval for the interior case by

\[
\text{CI}^{\text{interior}}_{1 - \alpha} = \left[ \hat{\theta} - \frac{c_{1 - \alpha}(T^*)}{\sqrt{n}}, \hat{\theta} + \frac{c_{\alpha}(T^*)}{\sqrt{n}} \right],
\]

where \( T \) is defined as \( T \). The reason why we use \( c_{\alpha}(T^*) \) and \( c_{1 - \alpha}(T^*) \) instead of \( c_{\alpha/2}(T^*) \) and \( c_{1 - \alpha/2}(T^*) \) is that when \( m_0 \in \mathcal{M} \), \( \hat{\theta}_0 < \hat{\theta}_0 \) 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 \( \hat{\theta}_0 = \bar{\theta}_0 = \theta_0 \), the asymptotic coverage of \( \text{CI}^{\text{interior}}_{1 - \alpha} \) is in general smaller than \( 1 - \alpha \). We consider instead the symmetric confidence interval

\[
\text{CI}^{\text{boundary}}_{1 - \alpha} = \left[ \hat{\theta} - \frac{c_{1 - \alpha}(T^*_s)}{\sqrt{n}}, \hat{\theta} + \frac{c_{1 - \alpha}(T^*_s)}{\sqrt{n}} \right],
\]

where \( \hat{\theta} = (\hat{\theta} + \bar{\theta})/2 \) and \( T_s = \sqrt{n} \left| \hat{\theta} - (\theta_0 + \bar{\theta}_0)/2 \right| \). When \( m_0 \in \partial \mathcal{M} \), \( \hat{\theta} \) is a consistent estimator of \( \theta_0 = (\theta_0 + \bar{\theta}_0)/2 \) and we show in the proof of Theorem 3.3 below that the bootstrap statistic \( T^*_s \) has the same distribution as \( T_s \). Thus, \( \text{CI}^{\text{boundary}}_{1 - \alpha} \) has an asymptotic coverage rate of \( 1 - \alpha \).

\[\text{Because } \partial \mathcal{M} \text{ is not convex, this projection may not be well defined. This is not an issue here. In this case, } \pi_{\partial \mathcal{M}}(\hat{m}) \text{ denotes any element in the set } \arg\min_m \|\hat{m} - m\|.\]
Finally, to obtain a confidence interval with a correct asymptotic coverage in all situations, we let

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

The idea is that we will eventually pick $CI_{1 - \alpha}^{\text{boundary}}$ when the true parameter is at the boundary, because $I_n \xrightarrow{P} 1$ in this case, and $CI_{1 - \alpha}^{\text{interior}}$ otherwise. The validity of this confidence interval, established in Theorem 3.3 below, relies on the following conditions.

**Assumption 3.2** There are unique solutions, $\bar{x}(m_0)$ and $\underline{x}(m_0)$, to Programs (3.2) and (3.3). Moreover, $m \mapsto q(\bar{x}(m_0), m)$ and $m \mapsto q(\underline{x}(m_0), m)$ are differentiable at $m_0$.

**Assumption 3.3** We either have

- $m_0 \in \hat{\mathcal{M}}$, $\underline{\theta}_0 < \bar{\theta}_0$ and $\bar{\theta}'(m_0, \cdot) \neq 0$, $\underline{\theta}'(m_0, \cdot) \neq 0$;

- $m_0 \in \partial \mathcal{M}$, with $\overline{C}_{m_0}$ a half space and the cdf of the asymptotic distribution of $T$ continuous at its $1 - \alpha$ quantile.

Assumption 3.2 reinforces Assumption 3.1. Taken together, they ensure that $\bar{\theta}$ and $\underline{\theta}$ are differentiable at $m_0$, which is necessary for the bootstrap to work. Note that the unicity requirement holds automatically when $m_0 \in \hat{\mathcal{M}}$, by Proposition 3.1. When $m_0 \in \hat{\mathcal{M}}$, we can still prove unicity when $g(F, m_{01}) = \int P(u, m_{01})du$, with $P(\cdot, m_{01})$ a polynomial of at most degree $K + 1$, using again Proposition 3.1. We conjecture that this remains true for many other $g$, but leave this issue for future research. The differentiability condition is satisfied by the Theil index. It also holds for the Duncan index as soon as $\underline{x}(m_0)$ does not contain $m_{01}$. When $m_0 \in \hat{\mathcal{M}}$, this case is likely to be an exception. When $m_0 \in \mathcal{M}$, this issue is more likely. However, differentiability of $\underline{\theta}$ can still be obtained, because $\underline{x}(m_0 + t)$ will also contain $m_{01} + t_1$, for $t = (t_1, \ldots, t_K)$ small enough. In other words, we can redefine $q$ and $x$ to exclude the coordinate of $x$ equal to $m_{01}$. The differentiability requirement is then satisfied for this modified function $q$.

Assumption 3.3 is rather mild. When $m_0 \in \hat{\mathcal{M}}$, the set of distributions $\mathcal{P}_{m_0}$ is infinite, so that $\underline{\theta}_0 < \bar{\theta}_0$ 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 3.3** Suppose that Assumptions 2.1 and 3.1-3.3 hold. Then, with probability one,

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

17
Theorem 3.3 shows that bootstrap confidence intervals are asymptotically valid in general. The crucial conditions for obtaining this result are the differentiability of the bounds, which holds here by Assumption 3.2, 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 & 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.

If \( \text{CI}_{1-\alpha} \) is asymptotically valid whether \( m_0 \) lies in the interior or at the boundary of \( \mathcal{M} \), it is not clear whether it is uniformly valid, i.e. whether it satisfies

\[
\lim_{n \to \infty} \inf_{F \in \mathcal{F}} \inf_{\theta_0 \in [\underline{\theta}_0, \overline{\theta}_0]} \Pr_{\theta_0}(\theta_0 \in \text{CI}_{1-\alpha}) \geq 1 - \alpha,
\]

where \( \mathcal{F} \) is a subset of \( \mathcal{D} \), the set of cdf on \([0, 1]\), \( F \) denotes the cdf of \( p \), \( \underline{\theta}_0 \) and \( \overline{\theta}_0 \) are the corresponding bounds and \( \Pr_{\theta_0} \) denotes the probability under \( F \). The confidence interval considered by Imbens & 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 & Han (2009), in a related context where the endpoints of the identification interval are estimated.

We consider another confidence interval that satisfies the uniformity requirement. Such a confidence interval is generally conservative, on the other hand. We restrict ourselves to sets \( \mathcal{F}_{u,v} \) defined by

\[
\mathcal{F}_{u,v} = \{ F \in \mathcal{D} / F(1 - u) - F(u) \geq v \}
\]

for \( 0 < u < 1/2 \) and \( v \in (0, 1) \). When considering small \( u \) and \( v \), \( \mathcal{F}_{u,v} \) essentially excludes Bernoulli distributions on \( p \), which makes sense since the asymptotic distributions of \( \widetilde{P} \) for such distributions are degenerated.

For any \( K \)-vector \( P \), let \( R(P) = \text{diag}(P) - PP' \). We consider the confidence region on \( P_0 \) with asymptotic level \( 1 - \alpha \) defined by

\[
I_{1-\alpha} = \{ P \in (0, 1)^K : (P - \widetilde{P})'R(P)^{-1}(P - \widetilde{P}) \leq \chi^2_k(1 - \alpha) \},
\]
where \( \chi^2_K(1-\alpha) \) is the \( 1-\alpha \) quantile of a \( \chi^2 \) distribution.\(^{11}\) Then define

\[
\text{CI}^2_{1-\alpha} = \left[ \inf_{m \in M: Qm \in I_{1-\alpha}} \theta(m), \sup_{m \in M: Qm \in I_{1-\alpha}} \bar{\theta}(m) \right].
\] (3.5)

**Theorem 3.4** Suppose that Assumption 2.1 holds. Then, for all \( 0 < u < 1/2 \) and \( v \in (0,1) \),

\[
\lim_{n \to \infty} \inf_{F \in F_{u,v}} \Pr_F(\theta_0 \in [\hat{\theta} - \frac{c_{1-\alpha}(T^* \sqrt{n}), \hat{\theta} - \frac{c_\alpha(T^*)}{\sqrt{n}}]) \geq 1 - \alpha.
\]

This theorem shows actually that \( \text{CI}^2_{1-\alpha} \) is uniformly valid for the whole set \( [\theta_0, \bar{\theta}_0] \), not only for \( \theta_0 \). This result is obtained under very mild assumptions. Even if these confidence intervals are conservative in general, our simulations suggest that they may still be very informative, especially when \( K \) is small.

When \( K \) is random and independent of \( p \), we can also define a bootstrap confidence interval. Letting \( \hat{m}_b = (\hat{m}_{b1}, ..., \hat{m}_{bk}) \) as before, we first draw \( K \) in its empirical distribution and then draw \( X \) conditional on \( K = k \) according to the vector of probabilities \( \hat{P}_b^k = Q^k(\hat{m}_{b1}, ..., \hat{m}_{bk})' \). The bootstrap confidence interval can then be obtained as previously.

Finally, the situation is similar when \( K \) is random but possibly dependent on \( p \). We first draw \( K \) in its empirical distribution and then draw \( X \) conditional on \( K = k \) according to the vector of probabilities \( \hat{P}_b^k = Q^k(\hat{m}_{b1}, ..., \hat{m}_{bk})' \), where \( \hat{m}_b^k \) is defined as \( \hat{m}_b \), for the subsample of units with \( K = k \). Then we compute the bootstrap bounds and bootstrap statistics using (3.4). Note that in this case, \( \theta_0 < \bar{\theta}_0 \) unless \( m^k \in \partial M \) for all \( k \). Thus, the confidence interval

\[
\text{CI}^3_{1-\alpha} = \left[ \hat{\theta} - \frac{c_{1-\alpha}(T^*)}{\sqrt{n}}, \hat{\theta} - \frac{c_\alpha(T^*)}{\sqrt{n}} \right],
\]

defined as in the interior case above, is valid as soon as \( m^k \not\in \partial M \) for one \( k \). If one is reluctant to make this assumption, replacing \( \alpha \) by \( \alpha/2 \) in this interval ensures that the interval is asymptotically conservative in all cases.

### 3.3 Test of the binomial mixture model

As mentioned previously, the binomial model is testable. In this subsection, we develop a test of this hypothesis that, as shown before, is equivalent to \( m_0 = Q^{-1}P_0 \in M \). The idea

\(^{11}\)An alternative would be to replace \( R(P) \) by \( \tilde{R} = \text{diag}(\tilde{P}) - \tilde{P}\tilde{P}' \) in \( I_{1-\alpha} \). There are two reasons for using \( R(P) \) instead of \( \tilde{R} \). First, in practice \( \tilde{R} \) is often singular because one of the component of \( \tilde{P} \) is zero. Second, it has been shown in the case of binomial models (not multinomial as here) that the finite sample performances of confidence intervals based on \( R(P) \) are far better than those using \( \tilde{R} \) (Blyth & Still, 1983).
is to approximate the distance between $m_0$ and $M$ by the one between $\tilde{m}$ and $M$. This leads to a consistent test because $\tilde{m}$ consistently estimates $m_0$ both under the null and the alternative hypotheses. Formally, let us consider the test statistic

$$S_n = \sqrt{n} \| \tilde{m} - \hat{m} \|.$$ 

To approximate the distribution of $S_n$ under the null, we use as previously the parametric bootstrap. In this case, no correction due to boundary effects is required. The important point is rather to define a bootstrap distribution that is drawn under the null hypothesis. We thus consider a bootstrap distribution of $X$ defined by the vector of probabilities $\hat{P}_2 = Q\tilde{m}$. Letting $S_{n2}^*$ denote the bootstrap counterpart of $S_n$, we define the critical region of the test by

$$C_n = \{S_n > c_{1-\alpha}^*(S_{n2}^*)\}.$$ 

Note that when $S_n = 0$, viz. if $\tilde{m} \in M$, we always accept $H_0$ and it is unnecessary to compute $c_{1-\alpha}^*(S_{n2}^*)$.

**Theorem 3.5** If $m_0 \in \mathcal{M}$, $\lim_{n \to \infty} \Pr(C_n) = 0$ with probability one. If $m_0 \in \partial \mathcal{M}$, $\overline{C}_{m_0}$ is a half space and $\alpha < 1/2$, $\lim_{n \to \infty} \Pr(C_n) = \alpha$ with probability one. Finally, if $m_0 \notin \mathcal{M}$, $\Pr(C_n) \to 1$ with probability one.

The theorem shows that the test has asymptotic size equal to $\alpha$, and is consistent. Of course, this does not mean that we reject the binomial mixture model whenever the true DGP does not satisfy this condition. It may happen that the true DGP is not a binomial mixture model but can be rationalized by such a model.

### 4 Monte Carlo simulations

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

We first study whether the constraint that $m_0$ belongs to $\mathcal{M}$ is binding in practice when estimating $m_0$. The data generating process is defined as in Subsection 2.5 ($\Phi^{-1}(p) \sim \mathcal{N}(-3.26, 2.13)$), and we estimate $\Pr(\tilde{m} \notin \mathcal{M})$ 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 $\mathcal{M}$ shrinks very quickly with $K$. For instance, with 200 units, the estimated probability (with 1,000 simulations) is one as soon as $K$ is 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 $\theta_0$, even though this true one is not reduced to a singleton.\footnote{This result is in line with the Monte Carlo simulations of Wood (1999), who focuses on the distribution of $p$ and estimates it with either a projection method or maximum likelihood. As here, his estimator is unique when $\hat{m} \notin M$. He shows that this holds generally for moderate to large $K$, even if $n$ is large.} 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 the estimated moments $\hat{m} \notin M$.

![Figure 2: Probability that the estimated moments $\hat{m} \notin M$.](image)

Note: each dot corresponds to 1,000 simulations with the DGP $\Phi^{-1}(p) \sim N(-3.26, 2.13)$.

Table 1 displays the properties of the estimated bounds and the confidence intervals $\text{CI}_{0.95}^1$ and $\text{CI}_{0.95}^2$ for different $n$ and $K$. We consider here both the Theil and Duncan indices,
and the data generating process is defined as before by $\Phi^{-1}(p) \sim \mathcal{N}(-3.26, 2.13)$. For this distribution, $T \simeq 0.543$ and $D \simeq 0.779$. As $K$ increases, the length of the estimated set decreases. When $K = 12$, the estimated set is almost always reduced to a singleton, even when $n = 10,000$.

CR$(\theta_0)$ denotes the coverage rate of the true parameter by the confidence intervals, while CR$([\theta_0, \theta_0])$ is the coverage rate of the whole identification interval. Overall, the estimator of the identification interval is quite precise even for small samples. In our setting, we only observe a significant bias on $\theta_0$, which however does not lead to a low coverage of the confidence intervals. Consistent with Figures 1 and 2, we see that even for $n = 10,000$, standard errors are far larger than the length of the identification region for $K \geq 9$. This means that for $K \geq 9$, uncertainty mostly stems from estimation, not from partial identification. As expected, the coverage rate of CI$^{1.0}_{0.95}$ is usually conservative, with a true coverage rate lying mostly between 0.93 and 1. This is logical, since with our DGP $\theta_0 \not\in \{\theta_0, \tilde{\theta}_0\}$, so that the asymptotic coverage is one. This is particularly true for $K = 3$, which is logical since in this case the theoretical bounds are quite remote from each other.

Because the bootstrap confidence intervals may not be uniformly consistent, we investigate how the DGP affects the coverage rate of these confidence interval. Table 2 displays these coverage rates with three alternative DGP, with $K$ fixed to 6. The first is a discrete one, where $p$ takes the values 0 and $1/3$ with probabilities 0.9 and 0.1 respectively. This DGP was chosen so that $m_0 \in \partial M$ for all $K \geq 3$, and the first two moments are close to those of our application. The second and third DGP are such that $\Phi^{-1}(p) \sim \mathcal{N}(-2.26, 2.13)$ and $\Phi^{-1}(p) \sim \mathcal{N}(-1.26, 2.13)$, leading to minority proportions higher than in our baseline specification ($E(p) \simeq 0.101$ and 0.238, respectively). As with our baseline specification, CI$^{1.0}_{0.95}$ is usually conservative. This was expected for the two continuous DGP, but we also observe this pattern with the discrete DGP, for which the asymptotic level is equal to 95%.

Monte Carlo simulations (not reported here) reveal that we get closer to 95% for very large $n$ ($n \geq 100,000$). For discrete DGP with higher minority proportions, we also observe levels that are closer to 95% for small or intermediate $n$.

---

13 This does not contradict the consistency result of Theorem 3.2. With $K = 12$ and $n = 10^8$, the estimated bounds are distinct in 97% of the draws and the average estimates are close from $\tilde{\theta}_0$ and $\theta_0$: $[0.525, 0.547]$ for the Theil and $[0.761, 0.798]$ for the Duncan.

14 The rather low coverage rate that we obtain for the Duncan when $K = 9$ and $n = 10,000$ may be due to numerical trouble in optimization. This is a case where the probability that the bootstrap estimator $\hat{m}_0^*$ belongs to $\hat{\mathcal{M}}$ is not negligible (around 3%). But contrary to the case where $K = 3$ or $K = 6$, the optimization needed to yield the bounds is typically difficult because $\mathcal{M}$ is very small. If optimization errors occur in the bootstrap bounds ($\hat{\mathcal{M}}$, $\hat{\mathcal{M}}$), this is likely to affect the confidence interval that we obtain.
Table 1: Performance of $\hat{\theta}$ and properties of CI$^{1}_{0.95}$ and CI$^{2}_{0.95}$.

<table>
<thead>
<tr>
<th>K</th>
<th>$[\theta_0, \bar{\theta}_0]$</th>
<th>n</th>
<th>$[E(\hat{\theta}), E(\bar{\theta})]$</th>
<th>$[E(\hat{\theta}), E(\bar{\theta})]$</th>
<th>CI$^{1}_{0.95}$</th>
<th>CI$^{2}_{0.95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Length CR($\theta_0$)</td>
<td>Length CR($\theta_0$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>[0.397, 0.633]</td>
<td>100</td>
<td>[0.475, 0.540] (0.231) (0.244)</td>
<td>0.832 0.968</td>
<td>0.912 0.952</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.421, 0.627] (0.091) (0.052)</td>
<td>0.46 0.986</td>
<td>0.568 0.985</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.398, 0.632] (0.024) (0.016)</td>
<td>0.302 1</td>
<td>0.347 0.986</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>[0.491, 0.573]</td>
<td>100</td>
<td>[0.527, 0.527] (0.136) (0.136)</td>
<td>0.559 0.94</td>
<td>0.794 0.977</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.525, 0.535] (0.064) (0.064)</td>
<td>0.279 0.982</td>
<td>0.4 0.999</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.503, 0.553] (0.034) (0.039)</td>
<td>0.207 0.935</td>
<td>0.251 0.999</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>[0.517, 0.558]</td>
<td>100</td>
<td>[0.528, 0.528] (0.112) (0.112)</td>
<td>0.433 0.937</td>
<td>0.733 0.985</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.536, 0.536] (0.047) (0.047)</td>
<td>0.195 0.971</td>
<td>0.322 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.531, 0.532] (0.029) (0.029)</td>
<td>0.12 0.991</td>
<td>0.175 1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>[0.524, 0.549]</td>
<td>100</td>
<td>[0.531, 0.531] (0.008) (0.008)</td>
<td>0.373 0.934</td>
<td>0.708 0.996</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.538, 0.538] (0.009) (0.009)</td>
<td>0.158 0.966</td>
<td>0.284 0.999</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.534, 0.534] (0.021) (0.021)</td>
<td>0.09 0.989</td>
<td>0.142 1</td>
<td></td>
</tr>
</tbody>
</table>

Duncan index

<table>
<thead>
<tr>
<th>K</th>
<th>$[\theta_0, \bar{\theta}_0]$</th>
<th>n</th>
<th>$[E(\hat{\theta}), E(\bar{\theta})]$</th>
<th>$[E(\hat{\theta}), E(\bar{\theta})]$</th>
<th>CI$^{1}_{0.95}$</th>
<th>CI$^{2}_{0.95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Length CR($\theta_0$)</td>
<td>Length CR($\theta_0$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>[0.563, 0.904]</td>
<td>100</td>
<td>[0.654, 0.743] (0.306) (0.310)</td>
<td>0.843 0.977</td>
<td>0.841 0.894</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.599, 0.899] (0.132) (0.029)</td>
<td>0.581 0.988</td>
<td>0.677 0.985</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.566, 0.904] (0.035) (0.008)</td>
<td>0.416 1</td>
<td>0.461 0.985</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>[0.713, 0.82]</td>
<td>100</td>
<td>[0.763, 0.764] (0.149) (0.149)</td>
<td>0.552 0.956</td>
<td>0.784 0.974</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.778, 0.79] (0.072) (0.067)</td>
<td>0.29 0.978</td>
<td>0.449 0.999</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.741, 0.804] (0.046) (0.038)</td>
<td>0.231 0.946</td>
<td>0.283 1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>[0.736, 0.801]</td>
<td>100</td>
<td>[0.773, 0.773] (0.109) (0.109)</td>
<td>0.411 0.938</td>
<td>0.714 0.982</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.791, 0.791] (0.042) (0.042)</td>
<td>0.165 0.916</td>
<td>0.345 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.792, 0.794] (0.019) (0.018)</td>
<td>0.077 0.847</td>
<td>0.196 1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>[0.758, 0.800]</td>
<td>100</td>
<td>[0.779, 0.779] (0.084) (0.084)</td>
<td>0.333 0.929</td>
<td>0.673 0.993</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.786, 0.786] (0.034) (0.034)</td>
<td>0.134 0.929</td>
<td>0.286 0.999</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.784, 0.784] (0.021) (0.021)</td>
<td>0.086 0.974</td>
<td>0.153 1</td>
<td></td>
</tr>
</tbody>
</table>

Note: for each $(n, K)$, simulations are based on 2,000 draws of samples. The distribution of $p$ is $\Phi^{-1}(p) \sim \mathcal{N}(\pm 3.26, 2.13)$, leading to $T \approx 0.543$ and $D \approx 0.779$. CR($\theta_0$) denotes the coverage rate of the confidence interval CI$^{1}_{0.95}$ (i.e. $\Pr(\theta_0 \in \text{CI}_{1-\alpha})$), and CR($[\theta_0, \bar{\theta}_0]$) the coverage rate of CI$^{2}_{0.95}$ (i.e. $\Pr([\theta_0, \bar{\theta}_0] \subset \text{CI}_{2-\alpha})$).
Table 2: Influence of the DGP on inference.

<table>
<thead>
<tr>
<th>DGP</th>
<th>([\hat{\theta}_0, \bar{\theta}_0])</th>
<th>(n)</th>
<th>([E(\hat{\theta}), E(\bar{\theta})])</th>
<th>(\text{CI}^1_{0.95})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Length</td>
</tr>
<tr>
<td><strong>Theil index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>[0.677, 0.677]</td>
<td>100</td>
<td>[0.656, 0.656]</td>
<td>(0.084) (0.084)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.667, 0.667]</td>
<td>(0.028) (0.028)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.673, 0.673]</td>
<td>(0.011) (0.011)</td>
</tr>
<tr>
<td>(\Phi^{-1}(p) \sim \mathcal{N}(-2.26, 2.13))</td>
<td>[0.485, 0.537]</td>
<td>100</td>
<td>[0.509, 0.511]</td>
<td>(0.082) (0.083)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.500, 0.518]</td>
<td>(0.038) (0.038)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.488, 0.534]</td>
<td>(0.013) (0.014)</td>
</tr>
<tr>
<td>(\Phi^{-1}(p) \sim \mathcal{N}(-1.26, 2.13))</td>
<td>[0.462, 0.495]</td>
<td>100</td>
<td>[0.472, 0.474]</td>
<td>(0.061) (0.061)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.469, 0.487]</td>
<td>(0.025) (0.024)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.463, 0.494]</td>
<td>(0.007) (0.007)</td>
</tr>
<tr>
<td><strong>Duncan index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>[0.931, 0.931]</td>
<td>100</td>
<td>[0.895, 0.895]</td>
<td>(0.075) (0.075)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.920, 0.920]</td>
<td>(0.019) (0.019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.928, 0.928]</td>
<td>(0.006) (0.006)</td>
</tr>
<tr>
<td>(\Phi^{-1}(p) \sim \mathcal{N}(-2.26, 2.13))</td>
<td>[0.644, 0.739]</td>
<td>100</td>
<td>[0.71, 0.712]</td>
<td>(0.076) (0.075)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.693, 0.721]</td>
<td>(0.043) (0.033)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.656, 0.736]</td>
<td>(0.021) (0.008)</td>
</tr>
<tr>
<td>(\Phi^{-1}(p) \sim \mathcal{N}(-1.26, 2.13))</td>
<td>[0.607, 0.666]</td>
<td>100</td>
<td>[0.64, 0.645]</td>
<td>(0.062) (0.061)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.622, 0.655]</td>
<td>(0.035) (0.030)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.609, 0.665]</td>
<td>(0.012) (0.008)</td>
</tr>
</tbody>
</table>

Note: for each DGP and each \(n\), simulations are based on 2,000 draws of samples. In the discrete DGP, \(p\) takes values 0 and 1/3, with probabilities 0.9 and 0.1, leading to \(T \simeq 0.677\) and \(D \simeq 0.931\). The two other DGP correspond to larger minority proportion than in the baseline specification (\(E(p) \simeq 0.101\) and \(0.238\), respectively). For the first, \(T \simeq 0.515\) and \(D \simeq 0.704\) while for the second, \(T \simeq 0.479\) and \(D \simeq 0.646\). In all cases, \(K = 6\). CR(\(\theta_0\)) denotes the coverage rate of the confidence interval \(\text{CI}^1_{0.95}\) (i.e. \(\text{Pr}(\theta_0 \in \text{CI}^1_{1-\alpha})\)).
5 An application to workplace segregation by nationality across French establishments

Understanding why and how employers make their hiring decisions and employees apply for jobs implies to be able to measure workplace segregation. For example, the issue of segregation is also related to employment and wage differentials across groups, either on sex or ethnic grounds. Early works focused on gender or race segregation across occupations or industries, see, e.g., Fields & Wolff (1991). Groshen (1991) is the first contribution to use the information available at the scale of establishments. Carrington & Troske (1995) use the 1983 CPS to compute Duncan indices for gender segregation across establishments, with a focus on small firms. Another strand of literature, which aims at linking skill dispersion with wage distribution, requires the computation of segregation indices. Kremer & Maskin (1996) and Kramarz et al. (1996) analyze, in the US and the French cases, how skill dispersion, measured 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.\textsuperscript{15} Carrington & Troske (1997) present new results on black/white segregation introducing a method to correct for small-unit bias. Hellerstein & Neumark (2008) use the 1990 Decennial Employer-Employee Database to measure workplace segregation by education, language and ethnicity. They compute adjusted indices using Carrington and Troske’s method.

In this section, we aim 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 estimate of the Theil index. As a matter of comparison, we also display the naive estimate and the ones proposed by Carrington & Troske (1997), Allen et al. (2009) and Rathelot (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 (1.8 million establishments). We restrict ourselves to the

\textsuperscript{15}Kremer & Maskin (1996) and Kramarz et al. (1996) interpret their segregation measure as a R-squared and suggest that using adjusted R-squared might be a way to deal with small-unit issues.
1.65 million establishments with less than 25 employees. Because citizenship is notably miscoded in the DADS, the minority group is defined as individuals born outside France and not French at their birth.

Table 3: Test of the binomial mixture model for $K \geq 8$.

<table>
<thead>
<tr>
<th>Unit size $K$</th>
<th>p-value of the bootstrap test</th>
<th>Unit size $K$</th>
<th>p-value of the bootstrap test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 8$</td>
<td>1</td>
<td>17</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>0.81</td>
<td>18</td>
<td>0.59</td>
</tr>
<tr>
<td>10</td>
<td>0.40</td>
<td>19</td>
<td>0.82</td>
</tr>
<tr>
<td>11</td>
<td>0.47</td>
<td>20</td>
<td>0.94</td>
</tr>
<tr>
<td>12</td>
<td>0.76</td>
<td>21</td>
<td>0.24</td>
</tr>
<tr>
<td>13</td>
<td>0.60</td>
<td>22</td>
<td>0.87</td>
</tr>
<tr>
<td>14</td>
<td>0.11</td>
<td>23</td>
<td>0.41</td>
</tr>
<tr>
<td>15</td>
<td>0.32</td>
<td>24</td>
<td>0.49</td>
</tr>
<tr>
<td>16</td>
<td>0.54</td>
<td>25</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Note: for $K \leq 8$, $\tilde{m} \in \mathcal{M}$, so that $S_n = 0$ and p-value = 1.

Before presenting our results, we first check that the binomial mixture model is not rejected in these data. For $K = 2\ldots 8$, we obtain that $\tilde{m} \in \mathcal{M}$, so the test is automatically accepted. For $K \geq 9$, $\tilde{m} \not\in \mathcal{M}$, but this may be expected even if $m_0 \in \mathcal{M}$ 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 3). We see this as evidence that the binomial mixture model is reasonable here.

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 $\tilde{m} \not\in \mathcal{M}$. 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 or 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 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). To test for this correlation, we consider the null hypothesis

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

where $\theta_0(K)$ is the true parameter corresponding to firms of size $K$. We test for this hypothesis at the nominal level $\alpha$ by rejecting the null when $\cap_{K \in \{5,10,15,20,25\}} \text{CI}_{\alpha}^1(\theta_0^K) = \emptyset$. It

\[\text{Pistaferri (1999) shows that, in Italy, smaller firms tend to use more often informal hiring channels. In a similar vein, Giuliano et al. (2009) show, for the US, that manager’s race affects the racial composition of new hires.}\]
is easy to see that by independence of the confidence intervals between the different sub-
samples, this test has asymptotically a true level smaller than \( \alpha \). The test is also consistent
against alternatives where \( \cap_{K \in \{5, 10, 15, 20, 25\}} \left[ \underline{\theta}_0^K, \bar{\theta}_0^K \right] = \emptyset \). For the Theil index, we do not
reject the null hypothesis at the 5\% level, though we reject it at 10\%. For the Duncan
index, we do not reject it at all standard levels. On the other hand, using the naive or
Allen et al. estimates, we reject at the 1\% level the hypothesis that segregation levels do
not depend on firm sizes.\(^{17}\) Thus, contrary to what is suggested by the naive and Allen et al.
estimates, we do not see much evidence that segregation levels depend on firm sizes in
France.

6 Conclusion

In this paper, we investigate what can be learned on segregation indices which are linear
functionals of \( F_p \) 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, as well as a test of the
binomial mixture model.

A first issue that we have not addressed here is nonlinear segregation indices, 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. Another interesting avenue of research, pioneered by Åslund & Skans
(2009), would be to study the dependence of segregation indices on unit characteristics,
such as sectors or geographical areas for firms.

\(^{17}\)The latter tests are based on the asymptotic normality of the naive and Allen et al. estimators. The
asymptotic variance matrices were computed by bootstrap.
References


A Comparison with other approaches

Formal characterization of the Allen et al. approach

One can show, in some circumstances, that $\theta_{ABW}$ reduces the order of the bias of $\theta_N$. Suppose that $g(F, m_{01}) = \int \gamma(x, m_{01}) dF(x)$, where $\gamma(., m_{01})$ is twice continuously differentiable and $\gamma(m_{01}, m_{01}) = 0$. Such restrictions are satisfied for the Theil index, for instance. Then, by a second order Taylor expansion and using decompositions of variances, we get

$$\theta_0 \approx \frac{\partial^2 \gamma}{\partial x^2}(m_{01}, m_{01}) \frac{V(p)}{2},$$
$$\theta_N \approx \frac{\partial^2 \gamma}{\partial x^2}(m_{01}, m_{01}) \frac{V\left(\frac{X}{K}\right)}{2},$$
$$\theta_N^* \approx \frac{\partial^2 \gamma}{\partial x^2}(m_{01}, m_{01}) \frac{V\left(\frac{X}{K}\right)}{2} \left[ V(p) + \frac{E(p(1-p))}{K} \right],$$

$$\theta^*_{N} \approx \frac{\partial^2 \gamma}{\partial x^2}(m_{01}, m_{01}) \frac{V\left(\frac{X}{K}\right)}{2} \left[ V(p) + \frac{E(p(1-p))}{K} \left( 2 - \frac{1}{K} \right) \right].$$

This suggests that the leading term in $\theta_N - \theta_0$, when $K \to \infty$, is $\frac{\partial^2 \gamma}{\partial x^2}(m_{01}, m_{01}) E(p(1-p))/K$. This term disappears in $\theta_{ABW} - \theta_0$, so that we expect the bias to be of smaller order, $1/K^{3/2}$ or $1/K^2$. However, as $\theta_{CT}$, $\theta_{ABW}$ may not lie in $[\theta_0, \theta_{0}]$, since it cannot be written in general as a segregation index $g(F_p, m_{01})$ with $p \in \mathcal{P}_{m_0}$. Note that by linearity of $g(., m_{01})$, $\theta_{ABW} = g\left(2F_{\bar{X}} - F_{X^*}, m_{01}\right)$ and the first moment of $2F_{\bar{X}} - F_{X^*}$ is $m_{01}$. But its second moment is $m_{02} + (m_{01} - m_{02})/K^2 \geq m_{02}$ with strict inequality in general, suggesting that $\theta_{ABW}$ is likely to overestimate the true segregation index.

Simulation results with $K$ random

Table 4 presents a comparison of the different approaches when the unit size is random, and uniform on $\{2, ..., 10\}$. The size of the identification interval crucially depends on whether one assumes $p \perp K$ or not. 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 4: Comparison between the sharp bounds, the naive approach and previous corrections, with a random sample size

<table>
<thead>
<tr>
<th>Method</th>
<th>Theil index</th>
<th>Duncan index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharp bounds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no independence</td>
<td>[0.46,0.59]</td>
<td>[0.65,0.84]</td>
</tr>
<tr>
<td>independence</td>
<td>[0.52,0.55]</td>
<td>[0.73,0.80]</td>
</tr>
<tr>
<td>Naive</td>
<td>0.73</td>
<td>0.92</td>
</tr>
<tr>
<td>Carrington-Troske</td>
<td>0.41</td>
<td>0.52</td>
</tr>
<tr>
<td>Allen et al.</td>
<td>0.67</td>
<td>0.91</td>
</tr>
<tr>
<td>Rathelot</td>
<td>0.55</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Note: $\Phi^{-1}(p) \sim N(-3.26, 2.13)$. With this DGP, the average size of the minority is $E(p) \simeq 0.033$, the Theil index is $T \simeq 0.543$ and the Duncan index is $D \simeq 0.779$.

B Additional Monte Carlo results

B.1 $K$ random

We consider the case where $K$ is random, and consider both inference based on the independence assumption $p \perp \perp K$ and inference without this assumption. We consider a DGP where $K$ takes the values 3, 6, 9, 12 with equal probabilities, $p \perp \perp K$ and $\Phi^{-1}(p) \sim N(-3.26, 2.13)$. Results are presented in Table 5. As expected, assuming independence allows one to shrink the length of confidence intervals. This is because it makes the identification problem negligible, but not thanks to a gain in accuracy in the estimated bounds. The gains due to the overidentifying equation (2.5) appear to be small, while the weighted average estimators based on (2.4) perform well in practice. Finally, the bootstrap confidence intervals behave well in both cases, except for small $n$ without the independence assumption where the coverage rate is well below 95%. This is due to the important bias of the bounds estimators for small sample sizes.
Table 5: Inference when $K$ is random, assuming or not $p \perp K$.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>$[\theta_0, \bar{\theta}_0]$</th>
<th>$n$</th>
<th>$[E(\hat{\theta}), E(\tilde{\theta})]$</th>
<th>Length</th>
<th>CR($\theta_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>($\sigma(\hat{\theta}))$ ($\sigma(\tilde{\theta})$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Theil index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>[0.522, 0.552]</td>
<td>100</td>
<td>[0.546, 0.546]</td>
<td>0.261</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.539, 0.539]</td>
<td>0.135</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.534, 0.534]</td>
<td>0.087</td>
<td>0.996</td>
</tr>
<tr>
<td>No independence</td>
<td>[0.481, 0.579]</td>
<td>100</td>
<td>[0.507, 0.524]</td>
<td>0.249</td>
<td>0.828</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.503, 0.557]</td>
<td>0.153</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.491, 0.562]</td>
<td>0.119</td>
<td>0.966</td>
</tr>
<tr>
<td><strong>Duncan index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>[0.745, 0.800]</td>
<td>100</td>
<td>[0.786, 0.786]</td>
<td>0.233</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.787, 0.787]</td>
<td>0.122</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.783, 0.783]</td>
<td>0.090</td>
<td>0.978</td>
</tr>
<tr>
<td>No independence</td>
<td>[0.693, 0.831]</td>
<td>100</td>
<td>[0.786, 0.812]</td>
<td>0.261</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>[0.743, 0.821]</td>
<td>0.189</td>
<td>0.951</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,000</td>
<td>[0.721, 0.821]</td>
<td>0.151</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: for each $n$, simulations are based on 2,000 draws of samples. $K$ takes values 3, 6, 9, 12, with probability $1/4$ for each. $p \perp K$ and $\Phi^{-1}(p) \sim \mathcal{N}(-3.26, 2.13)$, leading to $T \approx 0.543$ and $D \approx 0.779$. We use CI$_{0.95}^1$ when assuming independence, and CI$_{0.95}^3$ otherwise.

**B.2 Test of the binomial model**

Table 6 displays some elements about the performance of the bootstrap test of the binomial model proposed in the previous section. We use the same discrete distribution of $p$ as before. The test performs well in practice, with true levels close to the nominal one except for $K = 3$ where it appears to be conservative, especially for small or moderate $n$. 

34
Table 6: Tests of the binomial model: true levels of the bootstrap test (for a nominal level of 5%).

<table>
<thead>
<tr>
<th>$K$</th>
<th>$n$</th>
<th>True level</th>
<th>$K$</th>
<th>$n$</th>
<th>True level</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>100</td>
<td>0.008</td>
<td>9</td>
<td>100</td>
<td>0.042</td>
</tr>
<tr>
<td>3</td>
<td>1,000</td>
<td>0.016</td>
<td>9</td>
<td>1,000</td>
<td>0.053</td>
</tr>
<tr>
<td>3</td>
<td>10,000</td>
<td>0.037</td>
<td>9</td>
<td>10,000</td>
<td>0.062</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0.045</td>
<td>12</td>
<td>100</td>
<td>0.027</td>
</tr>
<tr>
<td>6</td>
<td>1,000</td>
<td>0.051</td>
<td>12</td>
<td>1,000</td>
<td>0.044</td>
</tr>
<tr>
<td>6</td>
<td>10,000</td>
<td>0.071</td>
<td>12</td>
<td>10,000</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Note: for each $(n, K)$, simulations are based on 2,000 draws. The distribution of $p$ takes values 0 and 1/3 with probability 0.9 and 0.1 respectively.

C Results from the application in the case in which $K$ is assumed to be random

We compute the bounds on the segregation indices for the whole set of firms. Results are displayed in Table 7. We first compute the minimum distance overidentification test that $p \perp K$ based on (2.5). Because of numerical issues, we restrict ourselves to firms of size 19 or less. We strongly reject the null hypothesis for these firms, with a test statistic $T \simeq 3.276$ while $T \sim \chi^2(170)$ under the null. Without the independence assumption, we estimate the bounds on the Theil (resp. on the Duncan) index to be $[0.440, 0.604]$ (resp. $[0.601, 0.812]$). The uncertainty is thus quite large, a result mostly driven by the lack of information on very small firms. Even if the estimated identified set is large, the Allen et al. estimate is outside this set for both indices. The same holds for the Carrington-Troske corrected Duncan. To improve our knowledge on segregation, we suppose that $p \perp K$ for $K \leq 8$. The overidentification test based on (2.5) is indeed accepted at all standard levels for firms of size 8 or less ($T \simeq 0.11$ with $T \sim \chi^2(27)$ under the null). We then compute the bounds combining both independence for $K \leq 8$ and non-independence for $K > 8$. We obtain in this case much more informative bounds, namely $[0.476, 0.508]$ for the Duncan and $[0.669, 0.728]$ for the Theil.
Table 7: Comparison between the sharp bounds, the naive approach and previous corrections on all firms.

<table>
<thead>
<tr>
<th>Method</th>
<th>Theil index Estimate</th>
<th>CI₀.₉₅</th>
<th>Duncan index Estimate</th>
<th>CI₀.₉₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharp bounds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>independence</td>
<td>[0.476, 0.508]</td>
<td>[0.472, 0.520]</td>
<td>[0.669, 0.728]</td>
<td>[0.655, 0.734]</td>
</tr>
<tr>
<td>no independence</td>
<td>[0.440, 0.604]</td>
<td>[0.434, 0.607]</td>
<td>[0.601, 0.812]</td>
<td>[0.596, 0.823]</td>
</tr>
<tr>
<td>Naive</td>
<td>0.739</td>
<td>[0.737, 0.74]</td>
<td>0.906</td>
<td>[0.903, 0.909]</td>
</tr>
<tr>
<td>Carrington-Troske</td>
<td>0.440</td>
<td>[0.437, 0.443]</td>
<td>0.546</td>
<td>[0.533, 0.556]</td>
</tr>
<tr>
<td>Allen et al.</td>
<td>0.675</td>
<td>[0.673, 0.677]</td>
<td>0.877</td>
<td>[0.874, 0.879]</td>
</tr>
<tr>
<td>Rathelot</td>
<td>0.501</td>
<td>[0.498, 0.503]</td>
<td>0.702</td>
<td>[0.700, 0.704]</td>
</tr>
</tbody>
</table>

Note: the bounds given under independence are obtained supposing only \( p \perp K \mid K \leq 8 \), not \( p \perp K \).

D Proofs

Proof of Theorem 2.1

We prove the result for the upper bound only, the reasoning being similar for the lower bound. Because any sequence of distributions on \([0, 1]\) is uniformly tight, the set \( \mathcal{D} \) of cdf on \([0, 1]\) is compact for the topology induced by the weak convergence. Because \( \mathcal{P}_{m₀} \) is closed for the weak convergence and \( \mathcal{P}_{m₀} \subset \mathcal{D} \), it is also compact. \( \mathcal{P}_{m₀} \) is also convex. Thus, by Krein-Milman theorem (see, Krein & Milman, 1940),

\[
\mathcal{P}_{m₀} = \overline{\text{co}}\left( \text{Ext}\left( \mathcal{P}_{m₀} \right) \right)
\]

where, for any set \( A \), \( \overline{\text{co}}(A) \) and \( \text{Ext}(A) \) denote the closure of the convex hull of \( A \) and the extremal elements of \( A \), respectively. Hence, by continuity of \( g(\cdot, m₀₁) \) with respect to the weak convergence,

\[
\overline{\theta}_0 = \max_{F \in \mathcal{P}_{m₀}} g(F, m₀₁) = \max_{F \in \overline{\text{co}}\left( \text{Ext}\left( \mathcal{P}_{m₀} \right) \right)} g(F, m₀₁) = \sup_{F \in \overline{\text{co}}\left( \text{Ext}\left( \mathcal{P}_{m₀} \right) \right)} g(F, m₀₁).
\]

If \( F = \sum_{i=1}^I \lambda_i F_i \) with \( \lambda_i \geq 0 \), \( \sum_{i=1}^I \lambda_i = 1 \) and \( F_i \in \text{Ext}\left( \mathcal{P}_{m₀}^{k} \right) \), we have, by linearity of \( g(\cdot, m₀₁) \),

\[
g(F, m₀₁) \leq \max \left( g(F₁, m₀₁), ..., g(F_I, m₀₁) \right).
\]

36
Therefore,
\[ \overline{\theta}_0 = \sup_{F \in \text{Ext}(\mathcal{P}_{m_0})} g(F, m_{01}). \]

To prove the result, it suffices to show that \( \text{Ext}(\mathcal{P}_{m_0}) \subset \mathcal{P}_{m_0}^{K+1} \). Suppose that \( F \in \text{Ext}(\mathcal{P}_{m_0}) \) has at least \( K + 2 \) support points. Then there exists \( (A_1, \ldots, A_{K+2}) \) disjoint subsets of \([0, 1]\) such that \( \int 1 \{ u \in A_j \} dF(u) > 0 \). By Theorem 1 of Douglas (1964), the functions \( x \mapsto x^0, \ldots, x \mapsto x^K \) are dense in \( L^1(F) \). Therefore, there exists \( (\gamma_{ij})_{i=1 \ldots K+2, j=1 \ldots K+1} \) such that
\[ 1_{A_i}(u) = \sum_{k=1}^{K+1} \gamma_{ik} u^{k-1} \quad F - \text{a.s.} \]

Let \( \nu_j = (\int_{A_j} 1 \{ u \in A_1 \} dF(u), \ldots, \int_{A_j} 1 \{ u \in A_{K+2} \} dF(u))' \), \( \alpha_j = (\int_{A_j} u^{0} dF(u), \ldots, \int_{A_j} u^{K} dF(u))' \) and \( \Gamma \) be the \((K + 2) \times (K + 1)\) matrix with typical \((i, j)\) term \( \gamma_{ij} \). We get \( \nu_j = \Gamma \alpha_j \) for \( j = 1, \ldots, K + 2 \). Because \( (\nu_1, \ldots, \nu_{K+2}) \) is a basis of \( \mathbb{R}^{K+2} \), this implies that the dimension of the range of \( \Gamma \) is at least \( K + 2 \), a contradiction. The result follows. 

**Proof of Proposition 2.3**

For any increasing and concave function \( u \), by Jensen’s inequality,
\[
E \left[ u \left( \frac{X}{K} \right) \right] = E \left[ E \left[ u \left( \frac{X}{K} \right) \mid p \right] \right] \\
\leq E \left[ u \left( E \left[ \frac{X}{p} \right] \right) \right] \\
\leq E[u(p)].
\]

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 \( F_p \in \mathcal{P}_{m_0} \) since such distributions rationalize the one of \( X/K \). Choosing a sequence \( (F_{n,p})_{n \in \mathbb{N}} \) in \( \mathcal{P}_{m_0} \) such that \( \lim_{n \to \infty} g(F_{n,p}, m_{01}) = \overline{\theta}_0 \), 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 \( E \left[ u \left( \frac{X}{K} \mid p \right) \right] < u \left( E \left[ \frac{X}{p} \right] \right) \) holds with a positive probability. As a result, \( E \left[ u \left( \frac{X}{K} \right) \right] < E[u(p)] \), and the result follows by strict monotonicity of \( g(\cdot, m_{01}) \).

**Proof of Theorem 3.2**

**Consistency** Because \( ||\hat{m} - m|| \leq ||\hat{m} - m|| \) and \( \hat{m} \) is consistent by the law of large numbers, \( \hat{m} \xrightarrow{P} m \). Moreover, \( (\hat{\theta}, \overline{\theta}) = (\hat{\theta}(\hat{m}), \overline{\theta}(\hat{m})) \). Therefore, by the continuous mapping theorem, it suffices to prove that \( \hat{\theta} \) and \( \overline{\theta} \) are continuous. We prove the result for the
upper bound only, the result being similar for the lower bound. First, remark that

\[ \tilde{\theta}(m) = \max_{F \in D} \tilde{g}(F, m) \quad \text{s.t.} \quad F \in G(m), \]

where \( \tilde{g}(F, m) = g(F, m_1) \) and \( G \) is the correspondence defined on \( \mathcal{M} \) to \( D \) by \( G(m) = \mathcal{P}_m \). To show continuity of \( \tilde{\theta} \), we check that the conditions of the Berge maximum theorem (see, e.g., Carter, 2001, Theorem 2.3) are satisfied. First, \( \tilde{g} \) is continuous as the composition of \( g \), which is continuous by Assumption 2.1, and the continuous function \( (F, m) \mapsto (F, m_1) \). Second, \( G \) is compact valued, since we proved above that \( \mathcal{P}_m \) is compact. Third, the domain and range of \( G \) are compact and the graph of \( G \) is \( \varphi^{-1}(\{0\}) \), with \( \varphi(m, F) = \int (x, \ldots, x^K)'dF - m \). Because \( \varphi \) is continuous, the graph of \( G \) is closed. As a consequence, \( G \) is upper hemicontinuous (see, e.g., Carter, 2001, Exercise 2.107).

Finally, we prove that \( G \) is lower hemicontinuous. We have to show that for any \((m, F), F \in G(m)\), and any sequence \( m_n \to m \) \((m_n \in \mathcal{M})\), there exists a subsequence \((m_{n_k})_k\) and \( F_{n_k} \in G(m_{n_k}) \) such that \( F_{n_k} \to_d F \). Let \( \bar{m}_n \in \partial \mathcal{M} \) and \( \lambda_n \in [0, 1] \) be such that

\[ m_n = \lambda_n m + (1 - \lambda_n) \bar{m}_n. \]

By Proposition 3.1, \( G \) is reduced to a singleton on \( \partial \mathcal{M} \). Let \( \{\tilde{F}_n\} = G(\bar{m}_n) \) and

\[ F_n = \lambda_n F + (1 - \lambda_n) \tilde{F}_n. \]

By construction, \( F_n \in G(m_n) \). Because \( \|m_n - m\| \to 0 \), we also have

\[ (1 - \lambda_n)\|m_n - \bar{m}_n\| \to 0. \]

If \( \liminf \|m - \bar{m}_n\| > 0 \), then \( \lambda_n \to 1 \), implying that \( F_n \to_d F \). On the other hand, if \( \liminf \|m - \bar{m}_n\| = 0 \), there exists a subsequence \((m_{n_k})_k\) such that \( \|m - m_{n_k}\| \to 0 \).

Because \( G \) is upper hemicontinuous and single-valued on \( \mathcal{M} \), it is continuous on \( \mathcal{M} \). As a result, \( F_{n_k} \to_d F \), implying also that \( F_{n_k} \to_d F \). Hence, in all cases, we have proved that there exists a subsequence \((m_{n_k})_k\) and \( F_{n_k} \in G(m_{n_k}) \) such that \( F_{n_k} \to_d F \). The result follows.

**Convergence in distribution**  The key point for the proof is the following lemma, which extends the delta method to functions with directional derivatives only.

**Lemma D.1** Suppose that \( f : \mathbb{R}^k \to \mathbb{R}^l \) is Lipschitz and Gateaux differentiable at \( \eta_0 \), with directional derivative \( f'(\eta_0, h) \) in direction \( h \). Let \( T_n \in \mathbb{R}^k \) be such that \( r_n(T_n - \eta_0) \xrightarrow{d} U \) for some sequence \( r_n \to \infty \) and a random element \( U \). Then \( r_n(f(T_n) - f(\eta_0)) \xrightarrow{d} f'(\eta_0, U) \) and \( r_n(f(T_n) - f(\eta_0)) - f'(\eta_0, r_n(T_n - \eta_0)) \xrightarrow{P} 0 \).

38
Proof: let $C_f$ denote the Lipschitz constant for $f$ and consider vectors $h_t \to h$ as $t \to 0$. We have

$$\left\| \frac{f(\eta_0 + th_t) - f(\eta_0) - f'(\eta_0, h)}{t} \right\| \leq \left\| \frac{f(\eta_0 + th) - f(\eta_0 + th_t) - f'(\eta_0, h_t)}{t} \right\| + \left\| \frac{f(\eta_0 + th) - f(\eta_0) - f'(\eta_0, h)}{t} \right\| \leq C_f \|h_t - h\| + o(1) = o(1),$$

where the first inequality follows by the triangular inequality, and the second by the fact that $f$ is Lipschitz and Gateaux differentiable at $\eta_0$. Similarly, for a given $\varepsilon > 0$, consider $h', h$ such that $\|h' - h\| < \varepsilon$. Fix also $t$ such that

$$\left\| \frac{f'(\eta_0, h') - f'(\eta_0, h)}{t} \right\| < \varepsilon, \left\| \frac{f'(\eta_0, h) - f(\eta_0 + th) - f(\eta_0)}{t} \right\| < \varepsilon.$$ 

By the triangular inequality,

$$\left\| f'(\eta_0, h') - f'(\eta_0, h) \right\| \leq \left\| f'(\eta_0, h') - f(\eta_0 + th') - f(\eta_0) \right\| + \left\| f(\eta_0 + th) - f(\eta_0 + th_t) \right\| + \left\| f(\eta_0 + th_t) - f(\eta_0 + th) \right\| \leq 3\varepsilon.$$

This implies that $f'(\eta_0, \cdot)$ is continuous. The result follows by remarking that the proof of Theorem 20.8 in van der Vaart (2000) does not involve linearity of $f'(\eta_0, \cdot)$. □

Now let us prove Theorem 3.2. For any closed convex set $C$, let $\pi_C$ denotes the projection on $C$. We have $\hat{m} = \pi_M(\tilde{m})$. Moreover, $\pi_M$ admits directional derivative in direction $h$ equal to $\pi_{C_{m_0}}(h)$ (see, e.g., Hiriart-Urruty, 1982). $\pi_M$ is also Lipschitz, as a projection. Besides, by the central limit theorem, $\sqrt{n}(\hat{m} - m_0) \overset{d}{\to} Z$. Therefore, by Lemma D.1,

$$\sqrt{n}(\hat{m} - m_0) = \sqrt{n}(\pi_M(\tilde{m}) - \pi_M(m_0)) \overset{d}{\to} \pi_{C_{m_0}}(Z). \quad (D.1)$$

By Assumption 3.1 and Theorem 3 of Morand et al. (2009), $B = (\tilde{\theta}, \tilde{\theta})$ is Lipschitz. Moreover, by Assumption 3.1 and Theorem 13 of Morand et al. (2009), $B$ is also directionally differentiable. Applying once more Lemma D.1, we get

$$\sqrt{n}(B(\hat{m}) - B(m_0)) \overset{d}{\to} B'(m_0, \pi_{C_{m_0}}(Z)).$$

where $B'(m_0, h)$ denote the directional derivative of $B$ at $m_0$ in the direction $h$. The result follows by definition of $B$. □
Proof of Theorem 3.3

We first prove the following lemma.

**Lemma D.2** Suppose that Assumption 3.3 holds. Then, conditional on \( \hat{m}_b \) and with probability approaching one,

\[
\sqrt{n} (\hat{m}^* - \hat{m}_b) \xrightarrow{d} \pi_{\mathcal{C}_{m_0}} (Z).
\]  (D.2)

If, further, Assumptions 3.1 and 3.2 hold, we also have

\[
\begin{pmatrix} T^* \\ T^* \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \hat{\theta}'(m_0, \pi_{\mathcal{C}_{m_0}}(Z)) \\ \theta'(m_0, \pi_{\mathcal{C}_{m_0}}(Z)) \end{pmatrix}.
\]  (D.3)

**Proof:** we first show that conditional on \( \hat{m}_b \) and with probability approaching one,

\[
\sqrt{n} (\tilde{m}^* - \hat{m}_b) \xrightarrow{d} Z.
\]  (D.4)

Our bootstrap is equivalent to drawing \((X_1^*, ..., X_n^*)\) iid with

\[
(\Pr(X_i^* = 1), ..., \Pr(X_i^* = K))' = \tilde{P}_b.
\]

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

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

Besides,

\[
V(I(X_i^*)|\tilde{m}_b) = \text{diag}(\tilde{P}_b) - \tilde{P}_b \tilde{P}_b' \xrightarrow{P} \text{diag}(P_0) - P_0 P_0'.
\]

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

\[
\sqrt{n} \left( \tilde{P}^* - \tilde{P}_b \right) \xrightarrow{d} N(0, \text{diag}(P_0) - P_0 P_0').
\]

This implies (D.4) since \( \tilde{m}^* = Q^{-1} \tilde{P}^* \).

Now, suppose that \( m_0 \in \mathcal{M} \). Then with probability approaching one, \( \hat{m}_b \in \mathcal{M} \) and thus also \( \tilde{m}^* \in \mathcal{M} \). As a result, with probability approaching one, \( \hat{m}^* = \tilde{m}^* \). Moreover, \( \mathcal{C}_{m_0} = \mathbb{R}^K \) so that \( \pi_{\mathcal{C}_{m_0}}(Z) = Z \). Hence, (D.2) holds when \( m_0 \in \mathcal{M} \).

Next, suppose that \( m_0 \in \partial \mathcal{M} \). Let \( Z_n^* = \sqrt{n} (\tilde{m}^* - \hat{m}_b) \). By the continuous mapping theorem, \( \pi_{\mathcal{C}_{m_0}}(Z_n^*) \xrightarrow{d} \pi_{\mathcal{C}_{m_0}}(Z) \). Therefore, it suffices to prove that

\[
\sqrt{n} (\hat{m}^* - \hat{m}_b) - \pi_{\mathcal{C}_{m_0}}(Z_n^*) \xrightarrow{P} 0.
\]  (D.5)
For that purpose, remark that by the proof of Theorem 3.2,
\[
\sqrt{n} (\hat{m}^* - \hat{m}_b) = \sqrt{n} (\hat{m}^* - m_0) + \sqrt{n} (m_0 - \hat{m}_b) \\
= \sqrt{n} (\pi_M(\hat{m}^*) - \pi_M(m_0)) + \sqrt{n} (m_0 - \hat{m}_b) \\
= \pi_{\mathcal{C}_{m_0}} (\sqrt{n}(\hat{m}^* - m_0)) + \sqrt{n} (m_0 - \hat{m}_b) + o_P(1). \tag{D.6}
\]

By Assumption 3.3, the boundary \(\partial \mathcal{C}_{m_0}\) of \(\mathcal{C}_{m_0}\) is linear. Thus, it is the tangent space of \(\mathcal{M}\) at \(m_0\), and by definition,
\[
\|\hat{m}_b - \pi_{\partial \mathcal{C}_{m_0}} (\hat{m}_b)\| = o_P (\|\hat{m}_b - m_0\|).
\]

As a result, letting \(u_n = \pi_{\partial \mathcal{C}_{m_0}} (\sqrt{n}(m_0 - \hat{m}_b))\), we get
\[
\|\sqrt{n} (m_0 - \hat{m}_b) - u_n\| = \sqrt{n} \|\hat{m}_b - \pi_{\partial \mathcal{C}_{m_0}} (\hat{m}_b)\| \\
= \sqrt{n} o_P (\|\hat{m}_b - m\| + \|m - m_0\|) \\
= o_P (\sqrt{n} \|\hat{m}_b - m_0\|) = o_P(1), \tag{D.7}
\]

where the first equality stems from linearity of \(\pi_{\partial \mathcal{C}_{m_0}}\), the second from the triangular inequality and the third from \(\|\hat{m}_b - m\| = \min_{m \in \partial \mathcal{M}} \|m - \hat{m}\|\). Combining (D.6) and (D.7) yields
\[
\sqrt{n} (\hat{m}^* - \hat{m}_b) = \pi_{\mathcal{C}_{m_0}} (\sqrt{n}(\hat{m}^* - m_0)) + u_n + o_P(1). \tag{D.8}
\]

Now, remark that
\[
\pi_{\mathcal{C}_{m_0}} (h) = \mathbf{1} \{h \in \mathcal{C}_{m_0}\} + \pi_{\partial \mathcal{C}_{m_0}} (h) \mathbf{1} \{h \notin \mathcal{C}_{m_0}\},
\]

where \(\pi_{\partial \mathcal{C}_{m_0}}\) is the linear projection onto the tangent space \(\partial \mathcal{C}_{m_0}\) of \(\mathcal{C}_{m_0}\). Besides, for all \(h_1 \in \mathbb{R}^K\) and \(h_2 \in \partial \mathcal{C}_{m_0}\), \(h_1 + h_2 \in \mathcal{C}_{m_0}\) if and only if \(h_1 \in \mathcal{C}_{m_0}\). As a result, for all \(h_1 \in \mathbb{R}^K\) and \(h_2 \in \partial \mathcal{C}_{m_0}\),
\[
\pi_{\mathcal{C}_{m_0}} (h_1) + h_2 = \pi_{\mathcal{C}_{m_0}} (h_1 + h_2). \tag{D.9}
\]

Hence,
\[
\sqrt{n} (\hat{m}^* - \hat{m}_b) = \pi_{\mathcal{C}_{m_0}} (\sqrt{n}(\hat{m}^* - m_0) + u_n) + o_P(1) \\
= \pi_{\mathcal{C}_{m_0}} (Z^*_n) + o_P(1),
\]

where the first equality follows by (D.8), (D.9) and the fact that \(u_n \in \partial \mathcal{C}_{m_0}\), and the second by (D.7) and the fact that projections are continuous. (D.5) follows.
Finally, we show (D.3). By Assumption 3.1 and Theorem 13 of Morand et al. (2009), \( \tilde{\theta} \) admits a directional derivative in the direction \( h \in \mathbb{R}^K \) equal to

\[
\tilde{\theta} (m_0, h) = \max \left\{ q_m' (x, m_0, h) - \bar{\lambda}(x)' \frac{\partial r}{\partial m} (x, m_0) h, \ x \in \arg \max_{y \in S_{K+1}: r(y, m) \geq c} q(y, m) \right\}
\]

where \( q_m' (x, m_0, h) \) denotes the directional derivative of \( q(x, \cdot) \) at \( m_0 \) in the direction \( h \), \( \bar{\lambda}(x) \) is the optimal lagrange multiplier and \( r(x, m) = A(x)^{-1} (1, m')' \). Thus, by Assumption 3.2,

\[
\tilde{\theta}' (m_0, h) = q' (\bar{x}(m_0), m_0) h - \bar{\lambda}(\bar{x}(m_0))' \frac{\partial r}{\partial m} (\bar{x}(m_0), m_0) h.
\]

This implies that \( \tilde{\theta}' (m_0, h) \) is linear. Therefore, \( \tilde{\theta} \) is differentiable at \( m_0 \). The same holds for \( \theta \). (D.3) follows by applying the delta method for the bootstrap (see, e.g., van der Vaart, 2000, Theorem 23.9) \( \square \)

Now let us prove Theorem 3.3. Suppose first that \( m_0 \in \tilde{\mathcal{M}} \). Then \( I_n \xrightarrow{P} 0 \) and it suffices to show that

\[
\inf_{\theta_0 \in \mathcal{I} \in \mathcal{L}} \lim_{n \to \infty} \Pr (\theta_0 \in \mathcal{C}_{1-\alpha}^{\text{interior}}) = 1 - \alpha. \tag{D.10}
\]

Suppose first that \( \theta_0 = \tilde{\theta}_0 \). Then

\[
\Pr \left( \theta_0 \in \mathcal{C}_{1-\alpha}^{\text{interior}} \right) = \Pr \left( \tilde{T} \leq c_{1-\alpha} (T^*) \right.
\]

\[
\leq \Pr \left( \tilde{T} \leq c_{1-\alpha} (T^*) \right.
\]

\[
- \Pr \left( \tilde{T} \leq c_{1-\alpha} (T^*) \right) \left. + \sqrt{n} (\tilde{\theta}_0 - \theta_0) \geq c_\alpha (T^*) \right) \tag{D.11}
\]

Let \( P_1 \) and \( P_2 \) denote the two probability terms in (D.11). As mentioned in Lemma D.2, \( \theta \) is differentiable at \( m_0 \), with a nonzero gradient by Assumption 3.3. Besides, when \( m_0 \in \mathcal{M} \), \( \pi_{c_{m_0}} (Z) = Z \). Thus, by Theorem 3.2, the asymptotic distribution of \( T \) is normal with strictly positive variance. This distribution is therefore continuous at \( c_{1-\alpha} (T) \). By Lemma D.2 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,

\[
P_2 \leq \Pr \left( \tilde{T} + \sqrt{n} (\tilde{\theta}_0 - \theta_0) < c_\alpha (T^*) \right) \to 0,
\]

since \( c_\alpha (T^*) = O_P (1) \) and \( \sqrt{n} (\tilde{\theta}_0 - \theta_0) \to \infty \). As a result, with probability one,

\[
\Pr \left( \theta_0 \in \mathcal{C}_{1-\alpha}^{\text{interior}} \right) \to 1 - \alpha.
\]

The same holds when \( \theta_0 = \tilde{\theta}_0 \). Finally, if \( \theta_0 \in (\tilde{\theta}_0, \bar{\theta}_0) \),

\[
\Pr \left( \theta_0 \in \mathcal{C}_{1-\alpha}^{\text{interior}} \right) = \Pr \left( \tilde{T} + \sqrt{n} (\tilde{\theta}_0 - \theta_0) \leq c_{1-\alpha} (T^*) \right. \left. + \sqrt{n} (\bar{\theta}_0 - \theta_0) \geq c_\alpha (T^*) \right) \tag{D.11}
\]

42
Because \( T + \sqrt{n}(\bar{\theta} - \theta_0) \to -\infty \) and \( \bar{T} + \sqrt{n}(\bar{\theta} - \theta_0) \to +\infty \), this latter probability tends to one. Hence, (D.10) holds.

Now suppose that \( m_0 \in \partial \mathcal{M} \). Because \( \theta_0 = \bar{\theta}_0 = \theta_0 \), we have, by Theorem 3.2, \( \sqrt{n}(\hat{\theta} - \bar{\theta}) = O_P(1) \). Because \( k_n \to \infty \), \( I_n \overset{P}{\to} 1 \) and it suffices to show that with probability one,

\[
\lim_{n \to \infty} \Pr \left( \theta_0 \in \text{CI}_{1-\alpha}^{\text{boundary}} \right) = 1 - \alpha. \tag{D.12}
\]

We have

\[
\Pr \left( \theta_0 \in \text{CI}_{1-\alpha}^{\text{boundary}} \right) = \Pr \left( \sqrt{n} \left| \hat{\theta} - \frac{\theta_0 + \bar{\theta}_0}{2} \right| \leq c_{1-\alpha}(T_*)^* \right)
\]

where in the first equality we have used the definition of \( \text{CI}_{1-\alpha}^{\text{boundary}} \) and the fact that \( \theta_0 = \bar{\theta}_0 = \theta_0 \). By Lemma D.1, Assumption 3.3 and Theorem 1.2.1 of Politis et al. (1999) once more,

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

with probability one. The result follows.

**Proof of Theorem 3.4**

Introduce the function \( I(x) = (\mathbf{1}\{x = 1\}, ..., \mathbf{1}\{x = K\})' \) and the vectors \( U_i(P_0) = R(P_0)^{-1/2}(I(X_i) - P_0) \) (where \( R(P_0)^{-1/2} \) denotes a square root matrix of \( R(P_0)^{-1} \)) and \( \hat{U}(P_0) = \frac{1}{n} \sum_{i=1}^{n} U_i(P_0) \). Hereafter \( P_0 \) depends on the cdf \( F \) of \( p \) but we omit this dependency in the absence of ambiguity. Let \( S_{1-\alpha} \) denote the sphere of radius \( \sqrt{\chi_K^2(1-\alpha)} \) in \( \mathbb{R}^K \). By definition of \( \text{CI}_{1-\alpha}^2 \),

\[
\Pr(F(P_0 \in I_{1-\alpha}) = \Pr(F(\sqrt{n} \hat{U}(P_0) \in S_{1-\alpha})). \tag{D.13}
\]

We have \( E(U_i(P_0)) = 0 \) and \( V(U_i(P_0)) = I_K \), the identity matrix of size \( K \). As a result, by the multivariate Berry-Esseen bound (see, e.g., Esseen, 1945, p.92),

\[
| \Pr(F(\sqrt{n} \hat{U}(P_0) \in S_{1-\alpha}) - (1 - \alpha) | \leq C(K) \frac{E_F(\sum_{j=1}^{K} U_{ij}(P_0)^4)}{n^{K/(K+1)}}, \tag{D.14}
\]

where \( Z \sim \mathcal{N}(0, I_K) \), \( C(K) \) is a constant independent of the distribution of \( X \) (and thus \( p \)) and \( U_{ij}(P_0) \) is the \( j \)th component of \( U_1(P_0) \).

We have

\[
E_F \left( \sum_{j=1}^{K} U_{ij}(P_0)^4 \right) \leq E \left[ \left( \sum_{j=1}^{K} U_{ij}(P_0)^2 \right)^2 \right] = E \left\{ \| U_1(P_0) \|^4 \right\}.
\]
Now, \[ \|U_1(P_0)\|^2 \leq \frac{\|I(X_1) - P_0\|^2}{\delta(P_0)} \leq \frac{K + 1}{\delta(P_0)}, \]
where \(\delta(P_0)\) is the smallest eigenvalue of \(R(P_0)\). Some algebra show that \(\delta(P_0) = \min_{i=1}^{K} P_{0i}\). Thus,

\[
E_F \left( \sum_{j=1}^{K} U_{ij}(P_0)^4 \right) \leq \frac{(K + 1)^2}{\min_{i} P_{0i}^2}.
\]

Now, for all \(i = 1, ..., K\), and all \(F \in \mathcal{F}_{u,v}\),

\[
P_{0i} = E_F [\Pr_F(X = i|p)] = C_i^K \int_0^1 p^i (1 - p)^{K-i} dF(p) \geq \int_u^1 p^i (1 - p)^{K-i} dF(p) = v^K \int_u^{1-u} \left(\frac{p}{u}\right)^i \left(\frac{1-p}{u}\right)^{K-i} dF(p) \geq u^K v,
\]

where the last inequality follows from \(F \in \mathcal{F}_{u,v}\). As a result,

\[
\sup_{F \in \mathcal{F}_{u,v}} E_F \left( \sum_{j=1}^{K} U_{ij}(P_0)^4 \right) \leq \frac{(K + 1)^2}{u^2 v^2}.
\]  
(D.15)

Combining (D.13), (D.14) and (D.15), we obtain

\[
\lim_{n \to \infty} \sup_{F \in \mathcal{F}_{u,v}} |\Pr_F(P_0 \in I_{1-\alpha}) - (1 - \alpha)| = 0.
\]  
(D.16)

Finally, remark that

\[
P_0 \in I_{1-\alpha} \Rightarrow \theta(Q^{-1}P_0) \in \theta(Q^{-1}I_{1-\alpha}), \quad \bar{\theta}(Q^{-1}P_0) \in \bar{\theta}(Q^{-1}I_{1-\alpha})\]

\[
\Rightarrow [\theta_0F, \bar{\theta}_0F] \subset \left[ \inf_{m \in M: Q m \in I_{1-\alpha}} \theta(m), \sup_{m \in M: Q m \in I_{1-\alpha}} \bar{\theta}(m) \right].
\]

Thus, by definition of \(CI_{1-\alpha}^2\), \(\Pr_F([\theta_0F, \bar{\theta}_0F] \subset CI_{1-\alpha}^2) \geq \Pr_F(P_0 \in I_{1-\alpha})\). The result follows by (D.16) \(\square\)

**Proof of Theorem 3.5**

First, if \(m_0 \in \mathcal{O}\), we have

\[
\Pr(C_n) \leq 1 - \Pr(S_n = 0) = 1 - \Pr(\tilde{m} \in \mathcal{M}) \to 0.
\]
Now, suppose that \( m_0 \in \partial M \), \( \overline{C}_{m_0} \) is a half space and \( \alpha < 1/2 \). As previously, we let hereafter \( S_n^* \) denote the bootstrap counterpart of \( S_n \), with the underlying vector of probability \( \hat{P}_b \) for \( X \). Note that when \( \tilde{m} \notin M \), or equivalently \( S_n > 0 \), we have \( \hat{m}_b = \hat{m} \). As a result,

\[
S_{n2}^*|S_n > 0 \overset{d}{=} S_n^*|S_n > 0.
\] (D.17)

Thus,

\[
\Pr(C_n) = \Pr(S_n > c_{1-\alpha}^*(S_n^*)|S_n > 0) \Pr(S_n > 0) = \Pr(S_n > c_{1-\alpha}^*(S_n^*)|S_n > 0) \Pr(S_n > 0) = \Pr(S_n > c_{1-\alpha}^*(S_n^*)).
\] (D.18)

Remark that \( \hat{m} - \tilde{m} = [\pi_M - \text{Id}](\hat{m}) - [\pi_M - \text{Id}](m_0) \). Then, applying Lemma D.1 to \( \pi_M - \text{Id} \) and using (D.4), we get

\[
\sqrt{n}(\hat{m} - \tilde{m}) = \sqrt{n}([\pi_M - \text{Id}]\hat{m} - [\pi_M - \text{Id}](m_0)) \xrightarrow{d} [\pi_{\overline{C}_{m_0}} - \text{Id}](Z).
\]

\[
\sqrt{n}(\hat{m}^* - \tilde{m}^*) = \sqrt{n}([\pi_M - \text{Id}]\hat{m}^* - [\pi_M - \text{Id}](\hat{m}_b)) \xrightarrow{d} [\pi_{\overline{C}_{m_0}} - \text{Id}](Z).
\]

An application of the continuous mapping theorem to the function \( x \mapsto ||x|| \) shows that \( S_n \) and \( S_n^* \) have the same limit in distribution \( S \). Moreover, because \( \overline{C}_{m_0} \) is a half space, there exists \( u_0 \in \mathbb{R}^K \) such that

\[
S = \left\| \pi_{\overline{C}_{m_0}}(Z) - Z \right\| = \max(u_0'Z, 0).
\]

Moreover, \( \Pr(u'_0Z \leq 0) = 1/2 < 1 - \alpha \). Thus, the distribution of \( S \) is continuous at its quantile \( c_{1-\alpha}(S) \). As a result, by Theorem 1.2.1 of Politis et al. (1999), and with probability one,

\[
\Pr(S_n > c_{1-\alpha}^*(S_n^*)) \to \alpha.
\]

The second point of the theorem follows using (D.18).

Finally, suppose that \( m_0 \notin M \). By the triangular inequality,

\[
||\hat{m} - \tilde{m}|| \geq ||\hat{m} - m_0|| - ||m_0 - \tilde{m}|| \geq ||\pi_M(m_0) - m_0|| - ||m_0 - \tilde{m}||.
\]

The first term tends to a positive constant, while the second term tends to zero in probability. Thus, \( S_n \) tends to infinity. On the other hand, a similar reasoning as previously shows that the asymptotic distribution \( S_{n2}^* \) is the one of \( ||\pi_{\overline{C}_{\pi_M(m_0)}}(Z)|| \). In other words, \( c_{1-\alpha}(S_n^*) = O_P(1) \). This implies that \( \Pr(C_n) \to 1 \).