Identification in models of sorting with social externalities

Maximilian Kasy

Department of Economics, UC Berkeley

Introduction

Urban areas around the world show large degrees of socioeconomic and ethnic segregation across neighborhoods - why?

Two polar explanations:

- **Sorting along exogenous** neighborhood **characteristics** X: households have different willingness (ability) to pay for those.
- **Social externalities**: households choose their neighborhood based on location choices of other households, i.e. neighborhood composition M.

Reality: probably both, but to what extent? Under what conditions is the degree of social externalities identified?

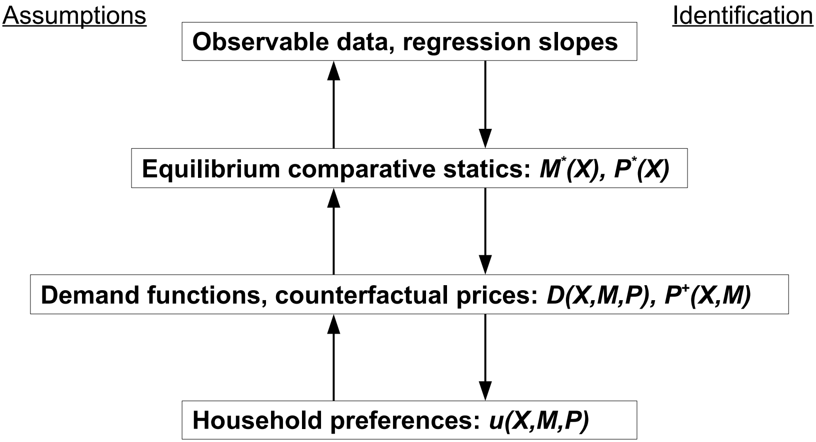
A fundamental identification problem arises in the model we will discuss:

- Neighborhood composition in equilibrium is functionally dependent on exogenous demand and supply determinants.
- $\Rightarrow$ The effects of composition on choices, i.e. **social externalities**, are **not identifiable**.
- Compare: “Simultaneity problem”, “Reflection problem”

Why should we care whether there are social externalities?

- Social externalities cause a **methodological problem** in the estimation of willingness to pay parameters that might inform policy,
- imply **multipliers** on policies affecting segregation,
- and may cause multiplicity of equilibria, **tipping**.

The big picture



Related problems

Other contexts with similar structure, sorting of:

- workers across firms
- students across schools
- customers across network providers
- faculty across universities
- spatial agglomeration and dispersion of firms

Some references - a very incomplete list

- **Sorting along amenities:** Tiebout (1956), Rosen (1974)
- **Sorting due to social externalities:** Schelling (1971), Becker and Murphy (2000), Nesheim (2001), Graham (2008)
- **Peer effects, identification:** Manski (1993), Moffitt (2004)
- **Empirical studies of sorting:** Black (1999), Chay and Greenstone (2005), Bayer, Ferreira, and McMillan (2007)
- **Search and matching:** Pissarides (2000), Wheaton (1990)

Roadmap

- Formal model
- Illustration of special case
- Negative identification results
- Positive identification results, based on:
    1. Subgroup shifters
    2. The spatial structure of cities
    3. The dynamic structure of prices in a search-model extension
- A LATE representation with identifiable weights
- Empirical application to US census data, focusing on Hispanic share in neighborhoods.
- Time permitting: A nonparametric test for multiple equilibria in the dynamics of neighborhood composition.

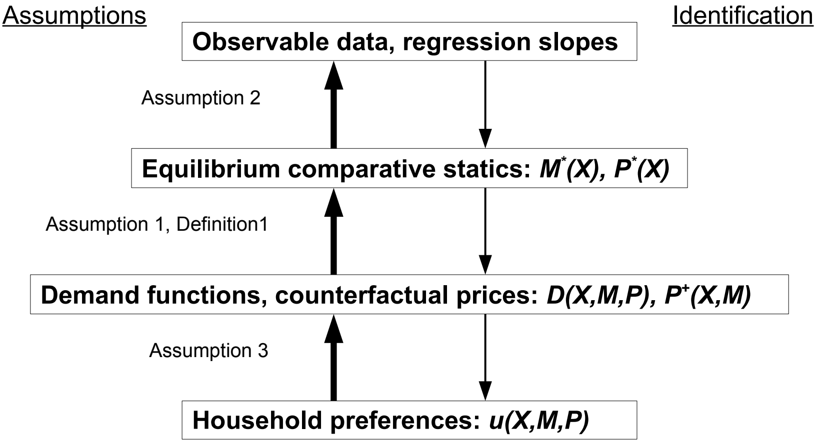
Baseline static model

3 assumptions:

1. **The local economy**: One neighborhood, demand and supply functions
   - Definition: Partial Sorting Equilibrium
   - Illustration: Two type case

   maps demand functions $\Rightarrow$ equilibrium schedules
2. **Observable data**: Many neighborhoods, data generated by equilibrium given exogenous factors
   maps equilibrium schedules $\Rightarrow$ observable data distribution
3. **Household utility maximization**:
   maps preferences $\Rightarrow$ demand functions

The big picture - assumptions



Assumption (1 - The local economy)

- *$\mathscr{C}$ types of households, $c = 1, \ldots, \mathscr{C}$.*
- *Neighborhood characterized by:*
  1. *Number of households of each type: $M = (M^1, \ldots, M^{\mathscr{C}})$*
  2. *Rental price: P*
  3. *Exogenous vector of all other location characteristics: X*
- *Demand for being at a neighborhood, for each type:*

$$\mathbf{D} = (\mathbf{D}^{\mathbf{1}}, \ldots, \mathbf{D}^{\mathscr{C}}) = \mathbf{D}(\mathbf{X}, \mathbf{M}, \mathbf{P})$$

  Total demand: $E = \sum_c D^c$.
- *Housing supply: $\mathbf{S}(\mathbf{P}, \mathbf{X})$.*

Definition (Partial Sorting Equilibrium)

A partial sorting equilibrium (M^*, P^*) *given* X *solves the* $C+1$ *equations*

$$D(X, M^*, P^*) = M^* \tag{1}$$

$$S(P^*, X) = \sum_c M^{*c} \tag{2}$$

Partial sorting equilibria given X*:* $(M^*(X), P^*(X))$

A special case for illustration

- Only $C = 2$ types.
- Both types have the same elasticity of demand with respect to prices and to the scale of the neighborhood.
- Define: $d = D^1/(D^1 + D^2)$, $m = M^1/(M^1 + M^2)$, and $E = D^1 + D^2$.

Under the above assumptions d is a function of m and X alone. This reduces the model to

$$d(m^*, X) = m^* \tag{3}$$

$$E(P^*, m^*, X) = S(P^*, X) \tag{4}$$

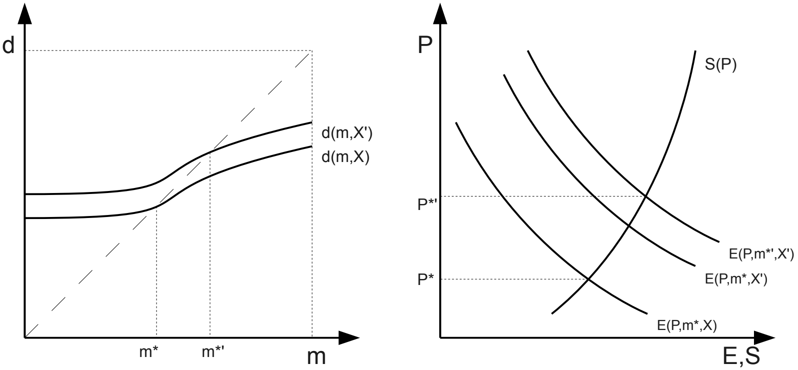


Figure: Comparative statics in the simplified $C = 2$ model

Assumption (2 - Observable data)

- *Repeated observations of* (X^1, M, P) *where* $X = (X^1, \epsilon)$ *for vectors* X^1 *and* ϵ.
- M **and** P **are in equilibrium given** X **for all observations**, *i.e.* $(M, P) \in (M^*(X), P^*(X))$.
- X *is continuously distributed on its support in* $\mathbb{R}^{dim(X)}$.
- *Full observability case:* $X = X^1$ *and* (M, P) *have full support on* $(M^*(X), P^*(X)) \Rightarrow (M^*(X), P^*(X))$ *is identified on support of* X.
- *Partial observability with exogenous variation case:* X^1 *is statistically independent of* ϵ *and the equilibrium selection mechanism.*

Assumption (3 - Household utility maximization)

- *Households characterized by:* $(u(X, M, P), u^o, c)$
- *Locate in the given neighborhood iff* $u(X, M, P) \geq u^o$.
- u^o *exogenously determined*
- *There is a continuum of households of total mass* M^{tot} *in the economy. The vector* (u, u_X, u_M, u_P, u^o)*, evaluated at any* (X, M, P)*, has a continuous joint distribution.*
- D^c *is the mass of households that want to locate in the given neighborhood,*

$$\mathbf{D^c} = \mathbf{M^{tot}} \cdot \mathbb{P}(\mathbf{u} \geq \mathbf{u^o}, \mathbf{c})$$

Similarly $E = M^{tot} \cdot \mathbb{P}(u \geq u^o)$.

The big picture - identification under full observability

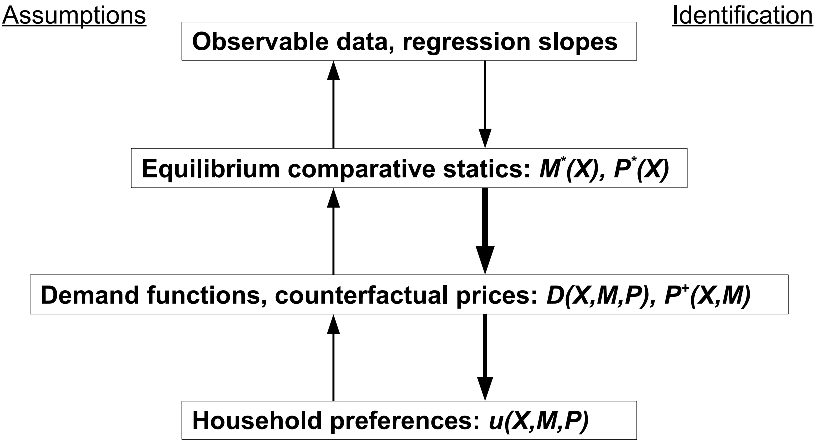


Table: COMPARISON TO PEER EFFECTS MODELS

Sorting with social externalities	Peer effects, as in Manski (1993) or Moffitt (2001)
Endogenous set of agents with fixed characteristics	Fixed set of agents with endogenous outcomes
Simultaneity problem: about identifying whether there are social externalities at all	Reflection problem / simultaneity: distinguishing endogenous from exogenous peer effects
Price mechanism allocating households to neighborhoods	-
Sorting is object of interest	Sorting is cause of identification problems, nuisance

Selected identification results

Notation:

- Subscripts $\Rightarrow$ partial derivatives
- Superscripts $\Rightarrow$ indices

Lemma (Price gradient as weighted average willingness to pay)

Make assumptions 1 and 3, and assume $S_P = S_X = 0$. Then

$$P_X^* = \widetilde{E}\left[-\frac{u_X + u_M M_X^*}{u_P} \middle| u = u^o \right],$$

where the expectation $\widetilde{E}$ is taken with respect to the density

$$f^{u_X, u_P | u - u^0}(u_X, u_P | 0) \cdot \frac{u_P}{E[u_P | u = u^0]}.$$

Proposition ((Non)identification)

Make assumptions 1 and 2, and consider the full observability case. Then:

- $D(X, M, P)$ *is not identified for* $(M, P) \notin (M^*(X), P^*(X))$.
- $D(X, M, P)$ *is identified on the joint support of* (X, M, P).

Corollary (Identification of slopes)

Linear combinations of the demand slopes are identified as

$$D_X + D_M M_X^* + D_P P_X^* = M_X^*. \tag{5}$$

No other linear combinations of (D_X, D_M, D_P) *are identified.*

Lemma (Spurious identification by functional form assumptions)

Make assumptions 1 and 2 and consider the full observability case, and assume that partial equilibrium is unique.

Fix an arbitrary $\mathscr{C} \times \mathscr{C}$ *matrix* A *and a* $\mathscr{C}$ *vector* B.

Then there **exists a just-identified model for** $D(X, M, P)$ **such that** $D_M \equiv A$ **and** $D_P \equiv B$ *for the unique* D *in the model such that* $D(X, M^*(X), P^*(X)) = M^*(X)$ *for all* X.

Positive identification results under full observability: Representing demand slopes in terms of equilibrium slopes.

Proposition (Subgroup identification)

Make assumptions 1 and consider the two type case. Assume that

- $D^1_{X^1} = 0$,
- *but* $D^2_{X^1} \neq 0$ *for some component* X^1 *of* X.

Then

$$D^1_m = \frac{1}{m^*_{X^1}} \left(M^{*1}_{X^1} - D^1_P P^*_{X^1} \right). \tag{6}$$

Assume additionally $D^1_{X^2} = D^1_{X^2} = 0$ *but* $S_{X^2} \neq 0$. *Then*

$$D^1_m = \frac{1}{m^*_{X^1}} \left(M^{*1}_{X^1} - \frac{M^{*1}_{X^2}}{P^*_{X^2}} P^*_{X^1} \right). \tag{7}$$

Spatial extension

Assumption (Cross neighborhood interactions)

- *There are $\mathscr{N}$ neighborhoods.*
- $\mathbf{G}$ *is a $\mathscr{N} \times \mathscr{N}$ matrix with non-negative entries, summing to one in each row, and with positive diagonal entries.*
- *Let* $\mathbf{m}$ *be the $\mathscr{N}$ vector of m for all neighborhoods,*
- $\widetilde{\mathbf{m}} = \mathbf{Gm}$ *the vector of* $\mathbf{G}$ *weighted averages of* $\mathbf{m}$*, and similarly* $\widetilde{\mathbf{X}} = \mathbf{GX}$.

Then, for each neighborhood, with $\widetilde{X}, \widetilde{m}$ being the neighborhood specific entries of the corresponding vectors,

$$d(\widetilde{m}, \widetilde{X}) = m \quad (8)$$

$$E(P^*, \widetilde{m}, \widetilde{X}) = S(P^*, X) \quad (9)$$

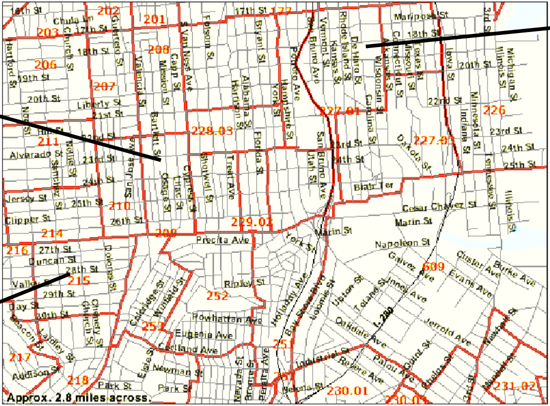
Illustration - census tracts in San Francisco

X^l excluded from d^i,
but X^l affects m^j and hence $\tilde{m}^i$

$$m^j = d(\tilde{m}^j, \tilde{X}^j)$$

$$\tilde{m}^i = \sum_j G^{ij} m^j$$

X^l



- Make assumption 4,
- assume $S_X = 0$ and $0 < d_{\widetilde{m}} < 1$,
- as well as $d_{\widetilde{X}} \neq 0$, for all neighborhoods.
- Fix two neighborhoods k and l.
- If the k,lth entry of $\mathbf{G}$ equals 0
- and there exists a power $j > 1$ of $\mathbf{G}$ such that the k,lth entry of $\mathbf{G}^j$ is not equal to zero, then:

Proposition (Spatial identification)

$$d_{\widetilde{m}}\left(\widetilde{m}^k, \widetilde{X}^k\right) = \frac{m^k_{X^l}}{\widetilde{m}^k_{X^l}} \tag{10}$$

and

$$D^c_{\widetilde{m}}\left(\widetilde{m}^k, \widetilde{X}^k, P^k\right) = \frac{1}{\widetilde{m}^k_{X^l}}\left(M^{*c,k}_{X^l} - D^{c,k}_P P^k_{X^l}\right) \tag{11}$$

Dynamic extension

Sketch of additional assumptions:

- **Continuous time**, X can change over time.
- **Search frictions**: Households trying to move find new place at rate λ. Landowners find tenants at rate μ.
- Therefore: Composition M changes continuously over time and only reacts with delay to shocks in X.
- **Match specific rental prices** P, landlords extract all surplus relative to outside option of breaking up.

Value functions of **households**, where $V = \max(V^s, V^{ns})$:

$$
\begin{aligned}
rV^s &= u(X, M, P) + \lambda(V^o - V) + \dot{V} && (12)\\
rV^{ns} &= u(X, M, P) + \dot{V} && (13)\\
(r+\lambda)V &= u(X, M, P) + \lambda \max(V^o, V) + \dot{V}. && (14)
\end{aligned}
$$

Value functions of **landowners**, where $W = \max(W^s, W^{ns})$:

$$
\begin{aligned}
rW^{ns} &= P + \dot{W} && (15)\\
rW^s &= P + \lambda(W^v - W) + \dot{W} && (16)\\
rW^v &= \mu(W^{new} - W^v) + \dot{W}^v && (17)
\end{aligned}
$$

Impulse response

Assume:

- $X = x$ before time 0, $X = x + \xi$ for a jump ξ after time 0.
- (u, V^o) is constant for all households.
- Average prices **before shock**: $P^b = \lim_{t \to 0^-} E[P]$
- **Short run**, after shock: $P^{sr} = \lim_{t \to 0^+} E[P]$
- **Long run**: $P^{lr} = \lim_{t \to \infty} E[P]$
- Long run composition: $M^{lr} = lim_{t \to \infty} M$

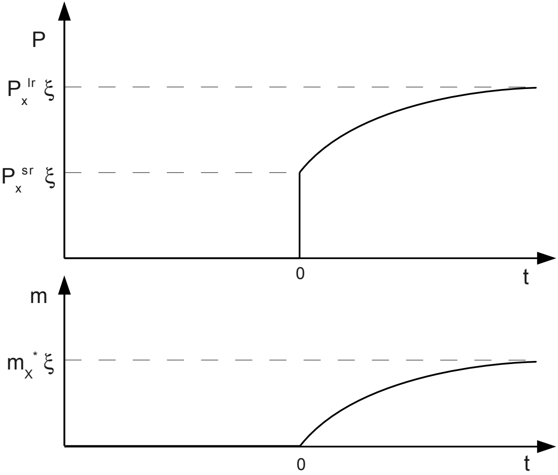


Figure: Dynamic response to shock in X

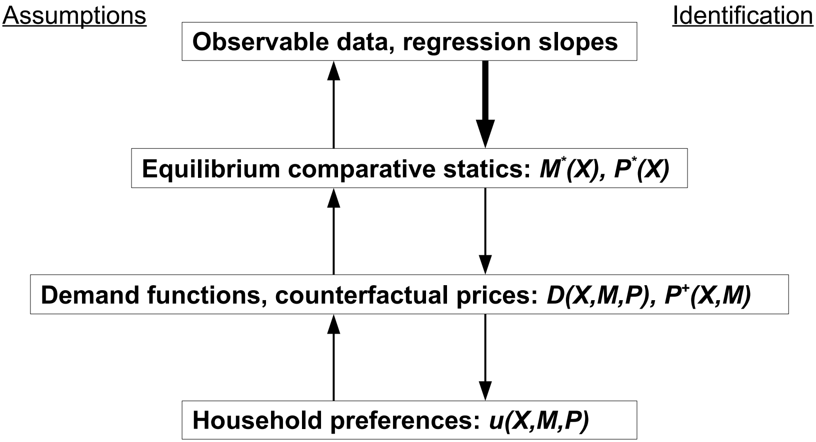
Proposition (Dynamic identification of hedonic slopes)

Under assumptions stated in the paper:

1. $P_\xi^{sr} = E\left[-\frac{u_X}{u_P}\right]$, $P_\xi^{lr} = E\left[-\frac{u_X + u_M M_\xi^{lr}}{u_P}\right]$ *and* $M_\xi^{lr} = M_X^*$.
2. *In the two type case,*
$$E\left[-\frac{u_m}{u_P}\right] = \frac{P_\xi^{lr} - P_\xi^{sr}}{m_\xi^{lr}}. \tag{18}$$
3. *More generally, for times* $t^2 > t^1 > 0$, *taking* P^{t^1}, P^{t^2} *as the time specific averages,*
$$E\left[-\frac{u_m}{u_P}\right] = \frac{P_\xi^{t^2} - P_\xi^{t^1}}{m_\xi^{t^2} - m_\xi^{t^1}}. \tag{19}$$

Completely analogous claims hold for any subgroup.

The big picture - identification of equilibrium schedules



Decomposing the LATE

Lemma (Crossectional IV with controls, random coefficient case)

Assume

$$Y^i = X^{1,i}\beta^{1,i} + X^{2,i}\beta^{2,i} + \epsilon \tag{20}$$

$$Z \perp (\beta, \epsilon) | X^2 \tag{21}$$

$$E[X^{2,i}\beta^{2,i} + \epsilon | X^2] \textit{ is linear in } X^2 \tag{22}$$

Denote $e = Z - E^*[Z|X^2]$. *Then*

$$\beta^{1,IV} = \frac{E[Ye]}{E[X^1 e]} = E\left[\frac{E[\beta^1 X^1 e | X^2]}{E[X^1 e]}\right] = E\left[\beta^{i,0} \cdot \omega\right]$$

for a weighting function

$$\omega = \frac{X^1 e}{E[X^1 e]}.$$

The point being

LATE-weights:
not in terms of latent first stage ("compliers," "always takers," ...)
but in terms of observables.

Suggestion:

- Plot the distribution of covariates reweighted by ω.
  $\Rightarrow$ This is the distribution of covariates for the population for which the IV coefficient describes the average partial effect.
- Plot nonparametric regressions of $\omega = \frac{X^1 \cdot e}{E[X^1 \cdot e]}$ and $Y \cdot e$ on components of X^2.
  $\Rightarrow$ These are the "conditional first stage" and the "conditional reduced form" (up to a constant).

Empirical application

- Neighborhood composition in cities in the United States.
- $C = 1$ is Hispanic, $C = 2$ non-Hispanic.
- Neighborhood Change Data Base (NCDB), aggregates data of the US census to the level of census tracts (on average ca. 4000 households/tract).
- Sample restricted to larger urban areas, outliers omitted.
- Imputed rents: share weighted average of observed rents and house values times estimated interest rate.

3 approaches - instruments used:

1. **Subgroup shifters**: predicted immigration, interacting national immigration and local population composition
2. **Spatial structure**: predicted immigration in neighborhoods 3km removed, conditional on predicted immigration locally
3. **Dynamic structure**: past composition change, conditional on current composition

Subgroup instrument - predicted immigration

Synthetic instrument dX^1 - interacting:

- national immigration from different source countries
- with local prior population from these source countries.

$$dX^1 = \frac{1}{M^1 + M^2} \sum_{\widetilde{c}} M^{\widetilde{c}} \cdot \frac{dM^{\widetilde{c},tot}}{M^{\widetilde{c},tot}} \tag{23}$$

- $\widetilde{c}$: Mexico, Puerto Rico and Cuba
- $M^{\widetilde{c}}$: initial population of type $\widetilde{c}$ in the neighborhood
- $M^{\widetilde{c},nat}$: total initial population of type $\widetilde{c}$
- $dM^{\widetilde{c},nat}$: total change of population of type $\widetilde{c}$

Assumption: dX^1 excluded from the demand of non-Hispanics.

Spatial instrument

Spatial instrument $\mathbf{dX}^{>3}$:

- average predicted change in Hispanic share, dX^1,
- in neighborhoods that are at least 3 km away
- but among the 15 closest neighborhoods

local average $\widetilde{\mathbf{m}}$:

- average Hispanic share
- in the given neighborhood and its 4 closest adjacent tracts

Assumption: $dX^{>3}$ excluded from local demand, $\widetilde{m}$ relevant composition for demand.

Dynamic instrument

Dynamic instrument X^L, in IV regressions of ΔP on Δm:
lagged Δm, controlling for m.

Assumption: Past changes in m uncorrelated with future changes in X, conditional on current m.

Table: INSTRUMENTAL VARIABLE ESTIMATES, DECADAL CHANGES IN THE 80S AND 90S

	first stage		IV regressions	
Instrument		log non-Hisp pop	log Hisp pop	log imputed rent
subgroup	0.146	-8.360	–	–
	(0.016)	(0.740)		
spatial	0.119	-6.251	3.437	-0.758
	(0.007)	(0.620)	(0.733)	(0.119)
dynamic	0.198	–	–	-0.516
	(0.011)			(0.049)

Notes: IV regressions, change in dependent variables on change in Hispanic share. Pooled data for the 80s and the 90s.

Controls for time x MSA fixed effects, and initial Hispanic share and its square (subgroup and dynamic instrument) or predicted immigration (spatial instrument).

Table: THEORETICAL INTERPRETATION OF THE ENTRIES OF TABLE 2

	first stage	IV regressions		
Instrument		log non-Hisp pop	log Hisp pop	log imputed rent
subgroup	$m^*_{X^I}$	$\frac{M^{*2}_{X^I}}{m^*_{X^I}} = \mathbf{D^2_m} + D^2_P \frac{P^*_{X^I}}{m^*_{X^I}}$	–	–
spatial	$\widetilde{m}_{X>3}$	$\frac{M^{*2}_{X>3}}{\widetilde{m}_{X>3}} = \mathbf{D^2_{\widetilde{m}}} + D^2_P \frac{P^*_{X>3}}{\widetilde{m}_{X>3}}$	$\frac{M^{*1}_{X>3}}{\widetilde{m}_{X>3}} = \mathbf{D^1_{\widetilde{m}}} + D^1_P \frac{P^*_{X>3}}{\widetilde{m}_{X>3}}$	$\frac{P_{X>3}}{\widetilde{m}_{X>3}} = \mathbf{P^+_{\widetilde{m}}} = \widetilde{E}\left[-\frac{\mathbf{u_{\widetilde{m}}}}{\mathbf{u_P}} \middle\vert u = u^o\right]$
dynamic	Δm_{X^L}	–	–	$\frac{\Delta P_{X^L}}{\Delta m_{X^L}} = E\left[-\frac{\mathbf{u_m}}{\mathbf{u_P}}\right]$

Interpretation

Assume rental elasticities D_P^2 between 0 and 2.
1% increase in Hispanic share $\Rightarrow$

- **Subgroup instrument:** 5 to 9% decline of non-Hispanic demand
- **Spatial instrument:**
  - 5 to 7% decline in non-Hispanic demand
  - 3 to 4% rise in Hispanic demand
  - 0.5% decline in housing costs
- **Dynamic instrument:** 0 to 0.5% decrease in average willingness to pay for home in the neighborhood.

Decomposition of the LATE

Recall

- e is the residual from regression of the instrument on the controls.
- $\omega = \frac{\Delta m \cdot e}{E[\Delta m \cdot e]}$, reweighting by ω gives the population over which the LATE averages.
- $E[\Delta m \cdot e|m]$ is the "conditional first stage," $E[\Delta M^2 \cdot e|m]$ is the "conditional reduced form."

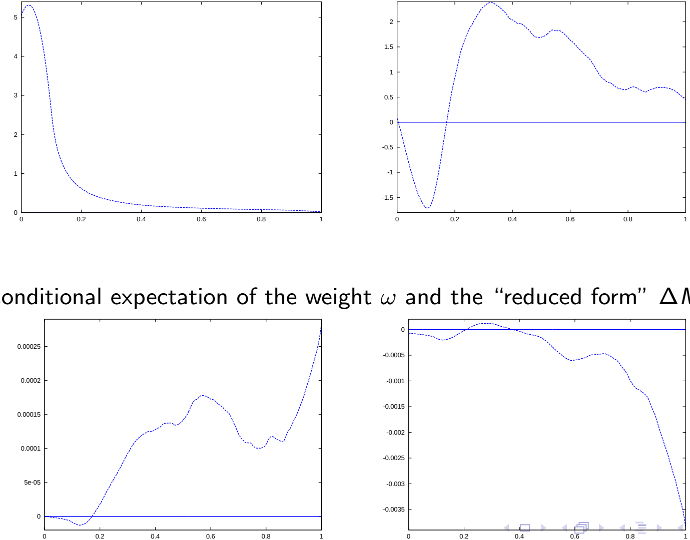
The following figure shows plots of kernel estimates of:

- The density of initial Hispanic share across neighborhoods, and this density reweighted by ω,
- $E[\omega|m]$, and $E[\Delta M^2 \cdot e|m]$.

Figure: DECOMPOSITION OF THE SUBGROUP IV ESTIMATE

Density and reweighted density of initial Hispanic share

Conditional expectation of the weight ω and the "reduced form" $\Delta M^2 \cdot e$



Bonus section: Dynamic model and multiple equilibria

Social externalities ($d_m \neq 0$)
$\Rightarrow$ possibly multiple equilibria of neighborhood composition
(i.e. solutions to the equation $d(m, X) = m$).

How can we test for the presence of such multiplicity?

Under the assumptions of the dynamic model:
Change of m from time 0 to 1 is given by

$$\Delta m = m^1 - m^0 \approx \kappa \cdot (d(m^0, X) - m^0). \tag{24}$$

Idea of test

- Identification:
  - Consider nonparametric quantile regressions of Δm on m.
  - The slope of regressions is upward biased.
  - This implies the number of roots is upward biased.
- Inference:
  - Estimate quantile regressions and their slopes using kernel method.
  - Plug the estimate into smoothed functional Z_σ approximating the number of roots.
  - Central result of "Nonparametric inference on the number of equilibria": Using non-standard asymptotics, the distribution of Z_σ converges to a normal, and we can perform inference on number of roots using t-statistics.

The following result shows: unstable equilibria of structural function
$\Rightarrow$ quantile regressions exhibit multiple roots.

Assumption (First order stochastic dominance)

$\mathbb{P}((d(m', X) - m') \leq Q|m)$ *is non-increasing as a function of* m*, holding* m' *constant.*

Proposition

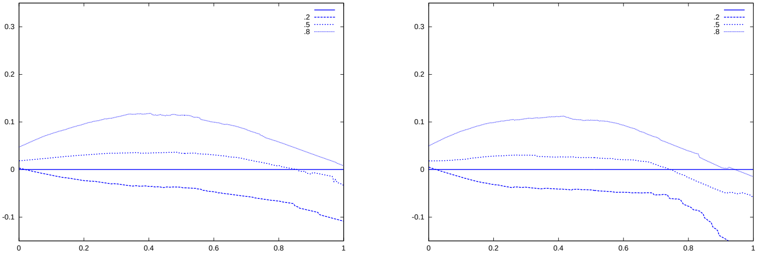
If $Q^{\Delta m|m}(\tau|m)$ *has only one root* m *for all* τ*, then the conditional average structural functions* $E\left[\kappa \cdot (d(m', X) - m')|d(X, m) = m, m\right]$*, as functions of* m'*, are "stable" at the roots* m*:*

$$E\left[\kappa \cdot (d_m - 1)|\Delta m = 0, m\right] \leq 0$$

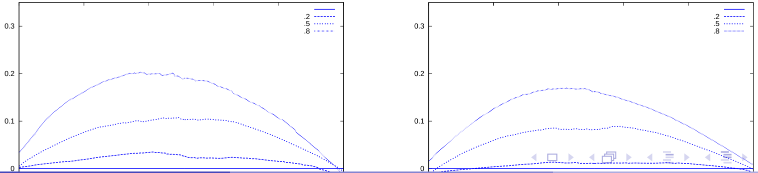
for all m*, where* $(0, m)$ *is in the support of* $(\Delta m, m)$*.*

Figure: Quantile regressions of change in Hispanic share on initial Hispanic share, 1980s and 1990s, .2, .5 and .8th quantile.

New York



Los Angeles



Inference on the number of roots

Assume we are interested in the number of roots $Z(g)$ of some function g on the range $\mathscr{M}$:

Definition

$$Z(g) := |\{m \in \mathscr{M} : g(m) = 0\}|$$

The inference procedure proposed is based on a smoothed version of Z, Z_σ:

Definition

$$Z_\sigma(g(.), g'(.)) := \int_0^1 L_\sigma(g(m))|g'(m)|dm$$

where L_σ is a Lipschitz continuous, positive symmetric kernel integrating to 1 with bandwidth σ and support $[-\sigma, \sigma]$

Let

- $g(m) = \mathrm{argmin}_d E_{\Delta m|m}[\rho(\Delta m - d)|m]$
- $(\hat{g}(m), \hat{g}'(m)) = \mathrm{argmin}_{a,b} \sum_k K_\tau(m_k - m)\rho(\Delta m_k - a - b(m_k - m))$
- $Z = Z(g)$ and $\hat{Z} := Z_\sigma(\hat{g}, \hat{g}')$

Theorem

Under assumptions stated in the paper, choosing $r_n = (n\tau^5)^{1/2}$, $n\tau \to \infty$, $\sigma \to 0$ *and* $\tau/\sigma^2 \to 0$ *there exist* $\mu > 0, V$ *such that*

$$\sqrt{\frac{\sigma}{\tau}}(\hat{Z} - \mu - Z) \to N(0, V)$$

for $\hat{Z} = Z_\sigma(\hat{g}, \hat{g}')$. *Both* μ *and* V *depend on the data generating process only via the asymptotic mean and variance of* $\hat{g}'$ *at the roots of* g.

Table: .95 confidence sets for $Z(g)$ for the largest MSAs by decade and quantile

MSA	**80s**			**90s**		
	$e = .2$	$e = .5$	$e = .8$	$e = .2$	$e = .5$	$e = .8$
New York	[1,1]	[0,0]	[0,0]	[0,0]	[1,1]	[0,0]
Los Angeles	[0,0]	[1,1]	[1,1]	[1,1]	[1,1]	[1,1]
Chicago	[1,1]	[1,1]	[0,0]	[0,0]	[0,0]	[0,0]
Houston	[0,1]	[0,0]	[0,0]	[0,1]	[1,1]	[0,0]
Phoenix	[1,3]	[0,0]	[0,0]	[1,1]	[0,0]	[0,0]
Philadelphia	[1,3]	[0,0]	[0,1]	[1,1]	[0,1]	[0,0]
San Antonio	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]
San Francisco	[1,1]	[1,1]	[0,0]	[2,2]	[1,1]	[0,0]

Summary - theoretical results

- Endogeneity of equilibrium composition at location leads to identification problem in models of sorting with social externalities.
- Source: The functional dependence of endogenous composition on exogenous demand shifters, both enter choices.
- Solutions have to "drive a wedge" between the two.
  1. *Subgroup shifters*: Assume some exogenous shifters do *not* enter choices of some subgroup.
  2. The *spatial structure* of cities: Assume externalities across adjacent locations, but not beyond $\Rightarrow$ propagation of composition shifts
  3. The *dynamic structure* of prices in a search-model extension: Prices react to shocks fast, composition adjusts with delay $\Rightarrow$ delayed price response to shocks identifies willingness to pay for composition.

Summary - empirical results

- These approaches were applied to neighborhoods in US cities, Hispanic share, aggregated census data.
- The approaches rely on problematic, but different, assumptions, yet they yield consistent estimates.
- Average willingness to pay for 1% increase in Hispanic share: around -0.5%
- Demand response to 1% increase in Hispanic share, holding prices constant:
  - 5 to 9% decline in non-Hispanic demand
  - 3 to 4% rise in Hispanic demand

Summary - decomposing the LATE, testing for multiple equilibria

- Linear IV estimates using controls can be decomposed as weighted averages of structural slopes with identifiable (!) weights.
- The instruments used here estimate the ATE for different subpopulations of neighborhoods in terms of initial Hispanic share.
- Strong social externalities imply multiple equilibria of neighborhood composition, which in turn imply multiple roots of quantile regressions of Δm on m.
- A test for the number of roots of such regressions allows to reject multiplicity of equilibria for almost all cities and decades considered.

Thanks for your time!