The Econometrics Journal

Econometrics Journal (2015), volume **00**, pp. 1–39.
doi: 10.1111/ectj.12043

Non-parametric inference on the number of equilibria

MAXIMILIAN KASY[†,‡]

[†]*Department of Economics, Harvard University, Littauer Center 200, 1805 Cambridge Street, Cambridge, MA 02138, USA.*
E-mail: maximiliankasy@fas.harvard.edu

[‡]*Institute for Advanced Studies, Stumpergasse 56, 1060 Vienna, Austria.*

First version received: June 2014; final version accepted: December 2014

Summary This paper proposes an estimator and develops an inference procedure for the number of roots of functions that are non-parametrically identified by conditional moment restrictions. It is shown that a smoothed plug-in estimator of the number of roots is superconsistent under i.i.d. asymptotics, but asymptotically normal under non-standard asymptotics. The smoothed estimator is furthermore asymptotically efficient relative to a simple plug-in estimator. The procedure proposed is used to construct confidence sets for the number of equilibria of static games of incomplete information and of stochastic difference equations. In an application to panel data on neighbourhood composition in the United States, no evidence of multiple equilibria is found.

Keywords: *Multiple equilibria, Non-parametric testing.*

1. INTRODUCTION

Some economic systems show large and persistent differences in outcomes even though the observable exogenous factors influencing these systems differ little.[1] One explanation for such persistent differences in outcomes is multiplicity of equilibria. If a system does have multiple equilibria, then temporary, large interventions might have a permanent effect, by shifting the equilibrium attained, while long-lasting, small interventions might not have a permanent effect.

Knowing the number of equilibria, and in particular whether there are multiple equilibria, is of interest in many economic contexts. Multiple equilibria and poverty traps are discussed by Dasgupta and Ray (1986), Azariadis and Stachurski (2005) and Bowles et al. (2006). Poverty traps can arise, for instance, if an individual's productivity is a function of their income and if wage income reflects productivity, as in models of efficiency wages. Productivity might depend on wages because nutrition and health are improving with income. If this feedback mechanism is strong enough, there might be multiple equilibria, and extreme poverty might be self-perpetuating. In that case, public investments in nutrition and health can permanently lift families out of poverty. Multiple equilibria and urban segregation are discussed by Becker and Murphy (2000) and Card et al. (2008). Urban segregation, along ethnic or sociodemographic dimensions, might arise because households' location choices reflect a preference over neighbourhood

[1] 'System' might refer to households, firms, urban neighbourhoods, national economies, etc.

composition. If this preference is strong enough, different compositions of a neighbourhood can be stable, given constant exogenous neighbourhood properties. Transition between different stable compositions might lead to rapid composition change, or 'tipping', as in the case of gentrification of a neighbourhood. Interest in such tipping behaviour motivated Card et al. (2008), and is the focus of the application discussed in Section 4 of this paper. Multiple equilibria and the market entry of firms are discussed by Bresnahan and Reiss (1991) and Berry (1992). Entering a market might only be profitable for a firm if its competitors do not enter that same market. As a consequence, different configurations of which firms serve which markets might be stable. In sociology, finally, multiple equilibria are of interest in the context of social norms. If the incentives to conform to prevailing behaviours are strong enough, different behavioural patterns might be stable norms (i.e. equilibria); see Young (2008). Transitions between such stable norms correspond to social change. One instance where this has been discussed is the assimilation of immigrant communities into the mainstream culture of a country.

This paper develops an estimator and an inference procedure for the number of equilibria of economic systems. It will be assumed that the equilibria of a system can be represented as solutions to the equation $g(x) = 0$. It will furthermore be assumed that g can be identified by some conditional moment restriction. The procedure proposed here provides confidence sets for the number $Z(g)$ of solutions to the equation $g(x) = 0$.

This procedure can be summarized as follows. In a first stage, g and its derivative g' are non-parametrically estimated. These first-stage estimates of g and g' are then plugged into a smooth functional Z_ρ, as defined in (2.4). We show that under standard i.i.d. asymptotics, and for the bandwidth parameter ρ small enough, the continuously distributed $Z_\rho(\widehat{g}, \widehat{g'})$ is equal to $Z(g)$ with probability converging to 1. A superconsistent estimator of $Z(g)$ can thus be formed by projecting $Z_\rho(\widehat{g}, \widehat{g'})$ on the closest integer.[2]

We then show that a rescaled version of $Z_\rho(\widehat{g}, \widehat{g'})$ converges to a normal distribution under a non-standard sequence of experiments. This non-standard sequence of experiments is constructed using increasing levels of noise and shrinking bandwidth as sample size increases. Under this same sequence of experiments, the bootstrap provides consistent estimates of the bias and standard deviation of $Z_\rho(\widehat{g}, \widehat{g'})$ relative to $Z(g)$. We can thus construct confidence sets for $Z(g)$ using t-tests. These confidence sets are sets of integers containing the true number of roots with a pre-specified asymptotic probability of $1 - \alpha$. An alternative to the procedure proposed here would be to use the simple plug-in estimator $Z(\widehat{g})$. This estimator just counts the roots of the first-stage estimate of g. We show, however, that the simple plug-in estimator is asymptotically inefficient relative to the smoothed estimator $Z_\rho(\widehat{g}, \widehat{g'})$ under the non-standard sequence of experiments considered.

Sections 3.4 and 3.5 discuss two general set-ups that allow us to translate the hypothesis of multiple equilibria into a hypothesis on the number of roots of some identifiable function g; these set-ups are static games of incomplete information and stochastic difference equations. Section 3.4 discusses a non-parametric model of static games of incomplete information, similar to the one analysed by Bajari et al. (2010).[3] Under the assumptions detailed in Section 3.4, we can non-parametrically identify the average best response functions (averaging over private information)

[2] An estimator is called superconsistent if it converges at a rate faster than the usual parametric rate, which equals the square root of the sample size.

[3] Other related papers from the recent literature include Aradillas-Lopez (2010), Lewbel and Tang (2011) and de Paula and Tang (2012). In contrast to these, we do not assume additively separable heterogeneity in latent payoffs. We can do this because we are only interested in response functions, not in latent utility. Note that our paper does not contribute to the literature discussing identification and estimation problems in games of complete information with multiple equilibria.

of the players in a static incomplete information game. This allows us to represent the Bayesian Nash equilibria of this game as roots of an estimable function. Section 3.4 discusses how to perform inference on the number of such Bayesian Nash equilibria.

Section 3.5 considers panel data of observations of some variable X, where X is generated by a general non-linear stochastic difference equation. This is motivated by the study of neighbourhood composition dynamics in Card et al. (2008). Section 3.5 argues that we can construct tests for the null hypothesis of equilibrium multiplicity of such non-linear difference equations by testing whether non-parametric quantile regressions of ΔX on X have multiple roots.

The rest of this paper is structured as follows. Section 2 presents the inference procedure and its asymptotic justification for the baseline case. Section 3 discusses generalizations, as well as identification and inference in static games of incomplete information and in stochastic difference equations. Section 4 applies the inference procedure to the data on neighbourhood composition studied by Card et al. (2008). In contrast to their results, no evidence of 'tipping' (equilibrium multiplicity) is found here. Section 5 concludes. Appendix A presents some Monte Carlo evidence. All proofs are relegated to Appendix B. Additional figures and tables are given in the online Appendix, which also contains a second application of the inference procedure to data on economic growth, similar to those discussed by Azariadis and Stachurski (2005), in their Section 4.1, and by Quah (1996).

2. INFERENCE IN THE BASELINE CASE

2.1. Set-up

Throughout this paper, the parameter of interest is the number of roots Z of some function g on a subset $\mathscr{X}$ of its support:

$$Z(g) := |\{x \in \mathscr{X} : g(x) = 0\}|. \tag{2.1}$$

Interest in this parameter is motivated by economic models in which the equilibria can be represented as roots of such a function g. Identification of the parameter $Z(g)$ follows from identification of g on $\mathscr{X}$. In this section, inference on $Z(g)$ is discussed for functions g with one-dimensional and compact domain and range. Throughout, the following assumption will be maintained.

ASSUMPTION 2.1. *(a) The observable data are i.i.d. draws of* (Y_i, X_i)*, where each draw has the same distribution as* (Y, X)*; (b) the set* $\mathscr{X}$ *is compact, and the density of* X *is bounded away from* 0 *on* $\mathscr{X}$*; (c) the function* g *is identified by a conditional moment restriction of the form*

$$g(x) = \operatorname{argmin}_y E_{Y|X}[m(Y - y)|X = x]; \tag{2.2}$$

(d) the function g *is continuously differentiable and generic in the sense of Definition 2.1.*

Examples of functions characterized by conditional moment restrictions as in (2.2) are conditional mean regressions, for which $m(\delta) = \delta^2$, and conditional qth quantile regressions, for which $m_q(\delta) = \delta \cdot (q - \mathbf{1}(\delta < 0))$.

DEFINITION 2.1 (GENERICITY). *A continuously differentiable function* g *is called generic if* $\{x : g(x) = 0$ *and* $g'(x) = 0\} = \varnothing$*, and if all roots of* g *are in the interior of* $\mathscr{X}$*.*

Genericity of g implies that g has only a finite number of roots.[4] Genericity in the sense of Definition 2.1 is commonly assumed in microeconomic theory; see the discussion in Mas-Colell et al. (1995, p. 593ff).

We propose the following inference procedure for the number of roots of g, $Z(g)$. First, estimate $g(\cdot)$ and $g'(\cdot)$ using local linear m-regression:

$$(\widehat{g}(x), \widehat{g'}(x)) = \text{argmin}_{a,b} \sum_i K_\tau(X_i - x) m(Y_i - a - b(X_i - x)). \tag{2.3}$$

Here, $K_\tau(\delta) = (1/\tau)K(\delta/\tau)$ for some (symmetric, positive) kernel function K integrating to one with bandwidth τ. Equation (2.3) is a sample analogue of (2.2), where a kernel weighted local average is replacing the conditional expectation. Next, calculate $\widehat{Z} = Z_\rho(\widehat{g}(\cdot), \widehat{g'}(\cdot))$, where Z_ρ is defined as

$$Z_\rho(g(\cdot), g'(\cdot)) := \int_{\mathscr{X}} L_\rho(g(x))|g'(x)|dx. \tag{2.4}$$

In this expression, $L_\rho(y) = (1/\rho)L(y/\rho)$ for a Lipschitz continuous, positive symmetric kernel L integrating to one with bandwidth 1 and support $[-1, 1]$. The intuition for this expression will be discussed in detail below. Estimate the variance and bias of $\widehat{Z}$ relative to Z using bootstrap. Finally, construct integer valued confidence sets for Z using t-statistics based on $\widehat{Z}$ and the bootstrapped variance and bias.

2.2. *Basic properties and consistency*

The rest of this section will motivate and justify this procedure. First, we see that $\widehat{Z}$ is a superconsistent estimator of Z, in the sense that $\alpha_n(\widehat{Z} - Z) \xrightarrow{p} 0$ for any diverging sequence $\alpha_n \to \infty$, under i.i.d. sampling and conditions to be stated. Then, we present the central result of this paper, which establishes asymptotic normality of $\widehat{Z}$ under a non-standard sequence of experiments. From this result, it follows that inference based on t-statistics, using bootstrapped standard errors and bias corrections, provides asymptotically valid confidence sets for Z. We also show that $\widehat{Z}$ is an efficient estimator relative to the simple plug-in estimator $Z(\widehat{g})$ under the non-standard asymptotic sequence.

We are mainly concerned with constructing confidence sets for Z, rather than a point estimator. A point estimator could be formed by projecting $\widehat{Z}$ on the closest integer. While $\widehat{Z}$ will be called an estimator of $Z(g)$, it should be kept in mind that its primary role is as an intermediate statistic in the construction of confidence sets.

The following proposition states that $Z(g) = Z_\rho(g)$ for generic g and ρ small enough. The two functionals only differ around non-generic g, or 'bifurcation points' (i.e. g where Z jumps). The functional Z_ρ is a smooth approximation of Z which varies continuously around such jumps.

PROPOSITION 2.1. *For g continuously differentiable and generic, if $\rho > 0$ is small enough, then*

$$Z_\rho(g(\cdot), g'(\cdot)) = Z(g(\cdot)).$$

[4] Suppose that g has an infinite number of roots in the compact set $\mathscr{X}$. Then, the set of x such that $g(x) = 0$ has an accumulation point in $\mathscr{X}$. At this accumulation point, genericity is violated.

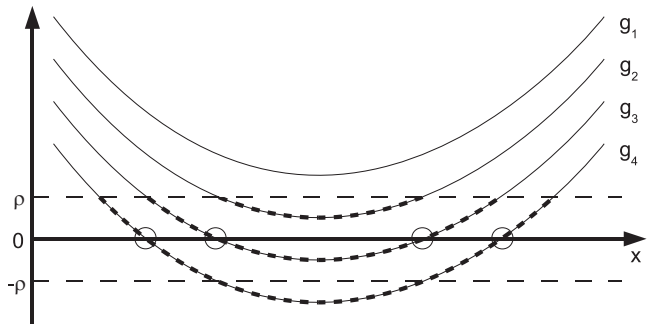

Figure 1. Z and Z_ρ.

The intuition underlying Proposition 2.1 is as follows. Given a generic function g, consider the subset of $\mathscr{X}$ where $L_\rho(g)$ is not zero. If ρ is small enough, this subset is partitioned into disjoint neighbourhoods of the roots of g, and g is monotonic in each of these neighbourhoods. A change of variables, setting $y = g(x)$, shows that the integral over each of these neighbourhoods equals one. Figure 1 illustrates the relationship between Z and Z_ρ. For the functions g depicted, $Z(g_1) = Z_\rho(g_1) = 0$, $Z(g_2) = 0 < Z_\rho(g_2) < 1$, $Z(g_3) = 2 > Z_\rho(g_3) > 1$ and $Z(g_4) = Z_\rho(g_4) = 2$. The two functionals are equal if g does not peak within the range $[-\rho, \rho]$, but if g does peak within the range $[-\rho, \rho]$, they are different and Z_ρ is not integer valued.

It is useful to equip the space of continuously differentiable functions on the compact set $\mathscr{X}$, $\mathscr{C}^1(\mathscr{X})$, with the following norm:

$$||g|| := \sup_{x\in\mathscr{X}} |g(x)| + \sup_{x\in\mathscr{X}} |g'(x)|. \tag{2.5}$$

This is the uniform first-order Sobolev norm on $\mathscr{C}^1(\mathscr{X})$. Given this norm, we have the following proposition.

PROPOSITION 2.2 (LOCAL CONSTANCY). *$Z(\cdot)$ is constant in a neighbourhood, with respect to the norm $||\cdot||$, of any generic function $g \in C^1$, and so is Z_ρ if ρ is small enough.*

Using a neighbourhood of g with respect to the sup norm in levels only, instead of $||\cdot||$, is not enough for the assertion of Proposition 2.2 to hold. For any function g_1 that has at least one root, we can find a function g_2 arbitrarily close to g_1 in the uniform sense, which has more roots than g_1, by adding a 'wiggle' around a root of g_1. Figure 2 illustrates. This figure shows two functions that are uniformly close in levels but not in derivatives, and which have different numbers of roots. However, if one additionally restricts the first derivative of g_2 to be uniformly close to the the derivative of g_1, additional wiggles are precluded around generic roots, because around these g_1 has a non-zero derivative. Because derivatives are 'harder' to estimate than levels, variation in the estimated derivatives dominates the asymptotic distribution of estimators for $Z(g)$, as will be shown. Proposition 2.2 immediately implies the following theorem as a corollary. This theorem states that the plug-in estimator $\widehat{Z} = Z_\rho(\widehat{g}(\cdot), \widehat{g'}(\cdot))$ converges to a degenerate limiting distribution at an 'infinite' rate, if $\widehat{g}$ converges with respect to the norm $||\cdot||$ (i.e. $\widehat{Z}$ is equal to the true number of roots with probability converging to 1).[5]

[5] The following theorem requires uniform convergence in probability of $(\widehat{g}, \widehat{g'})$ to (g, g'). Note that this is a slightly different condition from convergence of $\widehat{g}$ w.r.t. the norm $||\cdot||$ because $\widehat{g'}$ need not equal $\widehat{g}'$.

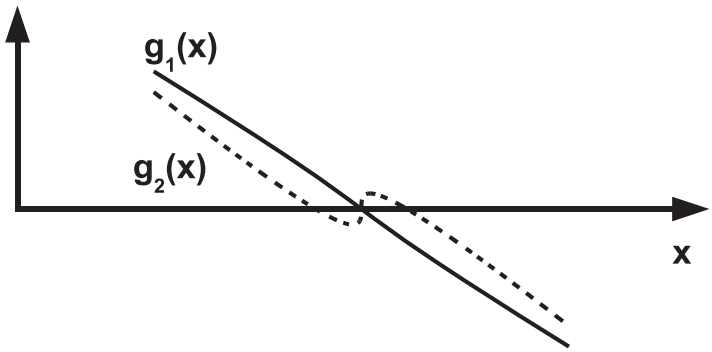

Figure 2. On the importance of wiggles.

THEOREM 2.1 (SUPERCONSISTENCY). *If $(\widehat{g}, \widehat{g'})$ converges uniformly in probability to (g, g'), if g is generic and if $\alpha_n \to \infty$ is some arbitrary diverging sequence, then*

$$\alpha_n(Z(\widehat{g}) - Z(g)) \xrightarrow{p} 0.$$

Furthermore, if ρ is small enough so that $Z_\rho(g, g') = Z(g)$ holds, then

$$\alpha_n(Z_\rho(\widehat{g}, \widehat{g'}) - Z(g)) \xrightarrow{p} 0.$$

This result implies that $\alpha_n(Z_\rho(\widehat{g}, \widehat{g'}) - Z(g)) \xrightarrow{p} 0$ if $\rho \to 0$ as $n \to \infty$.

2.3. Asymptotic normality and relative efficiency

We have shown our first claim, superconsistency of $\widehat{Z}$ given uniform convergence of $(\widehat{g}, \widehat{g'})$. Next, we show our second claim, asymptotic normality of $\widehat{Z}$ under a non-standard sequence of experiments. This section then concludes by formally stating the efficiency of $\widehat{Z}$ relative to the simple plug-in estimator $Z(\widehat{g})$. To further characterize the asymptotic distribution of $\widehat{Z}$, we need a suitable approximation for the distribution of the first-stage estimator $(\widehat{g}(\cdot), \widehat{g'}(\cdot))$. Kong et al. (2010) provide uniform Bahadur representations for local polynomial estimators of m-regressions. We state their result, for the special case of local linear m-regression, as an assumption.

ASSUMPTION 2.2 (BAHADUR EXPANSION). *The estimation error of the estimator $(\widehat{g}(x), \widehat{g'}(x))$ defined by (2.3) can be approximated by a local average as follows:*

$$\begin{aligned}(\widehat{g}(x), \widehat{g'}(x)) - (g(x), g'(x)) &= R - f_x^{-1}(x)s^{-1}(x)I_n(x) \\ &\times \frac{1}{n}\sum_i K_\tau(X_i - x)\phi(Y_i - g(x) - g'(x)(X_i - x))\Big(1, \frac{X_i - x}{\nu_2 \tau^2}\Big). \end{aligned} \tag{2.6}$$

Here, f_x is the density of X, $\nu_2 := \int K(x)x^2 dx$, $\phi := m'$ (in a piecewise derivative sense; m is assumed to be piecewise differentiable), $s(x) = \partial/(\partial y)E[\phi(Y - y)|X = x]|_{y=g(x)}$, $I_n(x)$ is a non-random matrix converging uniformly to the identity matrix, and $R = o_p((\widehat{g}(x), \widehat{g'}(x)) - (g(x), g'(x)))$ uniformly in x.

The crucial part of Assumption 2.2 is the assumption that the remainder R is asymptotically negligible relative to the linear (sample mean) component of $(\widehat{g}(x), \widehat{g'}(x)) - (g(x), g'(x))$. This assumption is only well defined in the context of a specific sequence of experiments.[6] In Theorem 2.2, this assumption will be understood to hold relative to the sequence of experiments defined in Assumption 2.3. In the case of qth quantile regression, $\phi(\delta) = q - \mathbf{1}(\delta < 0)$ and $s(x) = -f_{y|x}(g(x)|x)$. In the case of mean regression, $\phi(\delta) = -2\delta$ and $s(x) = -2$.

The asymptotic results in the remainder of this section depend on the availability of an expansion in the form of expansion (2.6) and the relative negligibility of the remainder, but not on any other specifics of local linear m-regression. This will allow for fairly straightforward generalizations of the baseline case considered here to the cases discussed in Section 3, as well as to other cases that are beyond the scope of this paper, once we have appropriate expansions for the first-stage estimators.

By Proposition 2.2, consistency of any plug-in estimator follows from uniform convergence of $(\widehat{g}(\cdot), \widehat{g'}(\cdot))$. Such uniform convergence follows from Assumption 2.2, combined with a Glivenko–Cantelli theorem on uniform convergence of averages, assuming i.i.d. draws from the joint distribution of (Y, X) as $n \to \infty$; see van der Vaart (1998), Chapter 19. Superconsistency of $\widehat{Z}$ therefore follows, which implies that standard i.i.d. asymptotics with rescaling of the estimator yield only degenerate distributional approximations. This is because Z_ρ and Z are constant in a C^1 neighbourhood of any generic g, even though they jump at bifurcation points (i.e. non-generic g). As a consequence, all terms in a functional Taylor expansion of Z_ρ, as a function of g, vanish, except for the remainder. The application of 'delta method' type arguments, as in Newey (1994), gives only the degenerate limit distribution.

In finite samples, however, the sampling variation of $\widehat{Z}$ is, in general, not negligible, as the simulations of Appendix A confirm, which makes the distributional approximation of the degenerate limit useless for inference. Asymptotic statistical theory approximates the finite sample distribution of interest by a limiting distribution of a sequence of experiments, of which our actual experiment is an element. The choice of sequence is to some extent arbitrary; the standard sequence where observations are i.i.d. draws from a distribution, which does not change as n increases, is just one possibility. In econometrics, non-standard asymptotics are used, for instance, in the literature on weak instruments; see, e.g. Staiger and Stock (1997), Imbens and Wooldridge (2007) and Andrews and Cheng (2012). In the present set-up, a non-degenerate distributional limit of $\widehat{Z}$ can only be obtained under a sequence of experiments, which yields a non-degenerate limiting distribution of the first-stage estimator $(\widehat{g}(\cdot), \widehat{g'}(\cdot))$.[7] We now consider asymptotics under such a sequence of experiments. The sequence we consider has increasing amounts of noise relative to signal as sample size increases.[8]

[6] Kong et al. (2010) provide regularity conditions under which

$$R = \left(1, \tfrac{1}{\tau}\right) O_p\left(\tfrac{\log(n)}{n\tau}\right)^{\lambda}$$

uniformly in X, for some $\lambda \in (0, 1)$ as $n \to \infty$ for stationary mixing processes.

[7] The approach of this paper, using local asymptotics, contrasts with the approach taken by most of the literature discussing inference on discrete valued parameters, testing and model selection. As argued by Choirat and Seri (2012), this literature has mostly focused on the use of large deviations asymptotics. The reason is that consistent estimators for discrete objects tend to converge at an exponential rate. Which type of asymptotics provides a more accurate approximation of finite sample distributions ultimately depends on the specific data-generating process; see Andrews and Cheng (2012). We should also mention the literature on testing for multimodality of densities (which is also based on i.i.d. asymptotics); see, e.g. Fischer et al. (1994).

[8] We could also define an equivalent sequence of experiments holding constant the amounts of noise and shrinking the signal.

ASSUMPTION 2.3. *Experiments are indexed by n, and for the nth experiment we observe* $(Y_{i,n}, X_{i,n})$ *for* $i = 1, \ldots, n$. *The observations* $(X_{i,n}, Y_{i,n})$ *are i.i.d. given n, and*

$$X_{i,n} \sim f_x(\cdot) \tag{2.7}$$

$$\gamma_{i,n}|X_{i,n} \sim f_{\gamma|X} \tag{2.8}$$

$$Y_{i,n} = g(X_{i,n}) + r_n\gamma_{i,n}, \tag{2.9}$$

where $\{r_n\}$ *is a real-valued sequence and*

$$0 = \operatorname{argmin}_a E[m(\gamma_{i,n} - a)|X_{i,n}] = \operatorname{argmin}_a E[m(r_n\gamma_{i,n} - a)|X_{i,n}].$$

The last equality requires the criterion function m to be scale neutral. This holds for quantiles and the mean, in particular. For a given sample size n, this is the same model as before. As n changes, the function g identified by (2.2) is held constant. If r_n grows in n, the estimation problem in this sequence of models becomes increasingly difficult relative to i.i.d. sampling. Note that (2.9) does not describe an additive structural model, which would allow us to predict counterfactual outcomes. Instead, $r_n\gamma_{i,n}$ is simply the statistical residual, given by the difference of Y and $g(X)$, which is also well defined for non-additive structural models.

Our next result, Theorem 2.2, assumes that the approximation of Assumption 2.2 holds under the non-standard sequence of experiments described by Assumption 2.3. Theorem 1 in Kong et al. (2010) implies that Assumption 2.2 holds under standard asymptotics and weak regularity conditions. Their result extends to our setting in a fairly straightforward way, however. This is most easily seen in the case of mean regression. We can write Y_i as a sum of two terms: (a) $g(X_{i,n}) + \gamma_{i,n}$; (b) $(r_n - 1) \cdot \gamma_{i,n}$. We can then apply the result of Kong et al. (2010) to local linear regression on $X_{i,n}$ of each of these terms separately. Both the Bahadur expansion and the local linear mean regression estimator are linear in Y. As a consequence, the remainder R for a regression of $Y_{i,n}$ on $X_{i,n}$ is given by the sum of the two remainders corresponding to regression of terms (a) and (b) on $X_{i,n}$. Whichever of the two Bahadur expansions corresponding to (a) and (b) dominates the asymptotic distribution is thereby guaranteed to be of larger order than the sum of the two remainder terms. A similar logic applies more generally, for instance to the case of local linear quantile regression; a complete proof is beyond the scope of the present paper.

By Corollary 2.1, a necessary condition for a non-degenerate limit of $\widehat{Z}$ is that $(\widehat{g}, \widehat{g'})$ converges to a non-degenerate limiting distribution. As is well known, and also follows from Assumption 2.2, $\widehat{g'}$ converges at a slower rate than $\widehat{g}$, so that asymptotically variation in $\widehat{g'}$ will dominate, namely by adding 'wiggles' around the actual roots. If $r_n = (n\tau^3)^{1/2}$ in the sequence of experiments defined in Assumption 2.3, $\widehat{g}$ converges uniformly in probability to g, whereas $\widehat{g'}$ converges pointwise to a non-degenerate limit. This is the basis for the following theorem.[9]

THEOREM 2.2 (ASYMPTOTIC NORMALITY). *Under Assumptions 2.1, 2.2 and 2.3, and if* $r_n = (n\tau^3)^{1/2}$, $n\tau \to \infty$, $\sqrt{\log(n)}\rho \to 0$ *and* $\tau/\rho^2 \to 0$, *then there exist* $\mu > 0$ *and* V *such that*

$$\sqrt{\frac{\rho}{\tau}}(\widehat{Z} - \mu - Z) \xrightarrow{d} N(0, V)$$

[9] The proof of Theorem 2.2 uses somewhat similar arguments as Horváth (1991) and Giné et al. (2003), who discuss the asymptotic distribution of the L^1 norm (L^p norm) of kernel density estimators.

for $\widehat{Z} = Z_\rho(\widehat{g}, \widehat{g'})$. *Both* μ *and* V *depend on the data-generating process only via the asymptotic mean and variance of* $\widehat{g'}$ *at the roots of* g, *which in turn depend upon* f_X, g', s *and* $Var(\phi|X)$ *evaluated at the roots of* g.

This theorem justifies the use of t-tests based on $\widehat{Z}$ for null hypotheses of the form $Z(g) = Z_\rho(g) = z_0$. The construction of a t-statistic requires a consistent estimator of V and an estimator of μ converging at a rate faster than $\sqrt{\rho/\tau}$. Based on the last part of Theorem 2.2, we can construct such estimators as follows. Any plug-in estimator that consistently estimates the (co)variances of $\widehat{g'}$ under the given sequence of experiments consistently estimates μ and V. One such plug-in estimator is standard bootstrap (i.e. resampling from the empirical distribution function). The Bahadur expansion in Assumption 2.2, which approximates $\widehat{g'}$ by sample averages, implies that the bootstrap gives a resampling distribution with the asymptotically correct covariance structure for $\widehat{g'}$. From this and Theorem 2.2, it then follows that the bootstrap gives consistent variance and bias estimates for Z_ρ, where the bias is estimated from the difference of the resampling estimates relative to $Z_\rho(\widehat{g})$. If sample size grows fast enough relative to $\sqrt{\rho/\tau}$ and τ, the asymptotic validity of a standard normal approximation for the pivot follows.

It would be interesting to develop distributional refinements for this statistic using higher-order bootstrapping, along the lines discussed by Horowitz (2001). However, higher-order bootstrapping might be very computationally demanding in the present case, in particular if criteria such as quantile regression are used to identify g.

Theorem 2.2 also implies that increasing the bandwidth parameter ρ reduces the variance without affecting the bias in the limiting normal distribution. Asymptotically, the difficulty in estimating Z is driven entirely by fluctuations in $\widehat{g'}$. These fluctuations lead both to upward bias and to variance in plug-in estimators. When ρ is larger, these fluctuations are averaged over a larger range of X, thereby reducing variance. Theorem 2.2 implies that Z_{ρ_1} is asymptotically inefficient relative to Z_{ρ_2} for $\rho_1 < \rho_2$. Furthermore, by Proposition 2.1, $Z(g) = \lim_{\rho \to 0} Z_\rho(g)$ for all generic g. If the relative inefficiency carries over to the limit as $\rho \to 0$, it follows that the simple plug-in estimator $Z(\widehat{g})$ is asymptotically inefficient relative to $\widehat{Z}$. Note, however, that this is only a heuristic argument. We cannot exchange the limits with respect to ρ and with respect to n to obtain the limit distribution of $Z(\widehat{g})$. The following theorem, which is fairly easy to show, states a formally correct version of this argument.

THEOREM 2.3 (ASYMPTOTIC INEFFICIENCY OF THE NAIVE PLUG-IN ESTIMATOR). *Consider the set-up of Theorem 2.2, and assume* $Z(g) > 0$. *Then, as* $n \to \infty$,

$$\liminf P(Z(\widehat{g}) > Z(g)) > 0$$

and

$$\mathrm{Var}\left(\sqrt{\frac{\rho}{\tau}} Z(\widehat{g})\right) \to \infty.$$

From this theorem, it follows in particular that tests based on $Z(\widehat{g})$ will, in general, not be consistent under the sequence of experiments considered (i.e. the probability of false acceptances does not go to zero). This stands in contrast to tests based on $\widehat{Z} = Z_\rho(\widehat{g}, \widehat{g'})$.

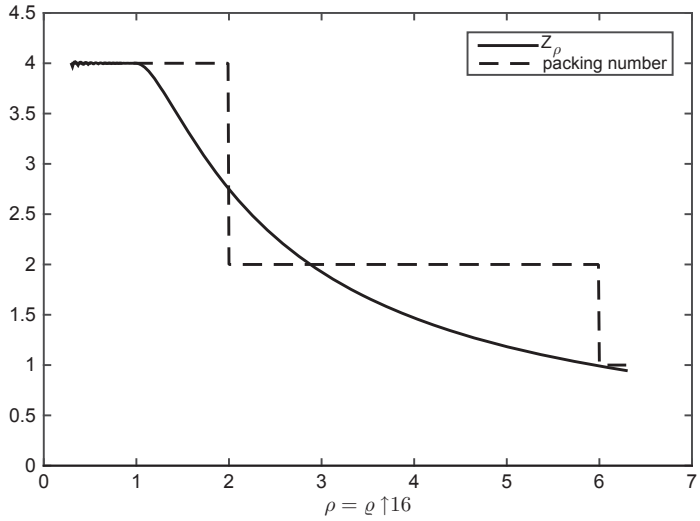

Figure 3. Z_ρ and the packing number.

2.4. Alternative approaches

The reader might wonder rightly whether there are alternative estimators that, like our $\widehat{Z}$, avoid the issues of the naive estimator (overestimating the number of roots, in particular), and that possibly beat Z_ρ in terms of some notion of relative efficiency.[10] One possible estimator that comes to mind is the ϱ-packing number of the set of roots of $\widehat{g}$, where $\varrho \to 0$ slowly. The packing number is the largest integer z such that there are z disjoint balls of radius ϱ centred at roots of $\widehat{g}$.

The packing number is in fact closely related to our estimator Z_ρ. For an appropriate scaling of ϱ, we can think of Z_ρ as smoothly interpolating the packing number. The following numerical illustration helps to make the point. Consider $g = \cos(x \cdot 4 \cdot \pi)$, and $x \in [0, 1]$. This function has four roots at a distance of $1/4$ from each other, and has a maximum absolute value of 1. For this function g, consider both $Z_\rho(g, g')$ and the packing number of g as a function of ρ (or ϱ). The result is plotted in Figure 3, which illustrates the relationship between $Z_\rho(g, g')$ and the packing number of the set of roots of g for the function $g = \cos(x \cdot 4 \cdot \pi)$, by plotting both as a function of bandwidth. For comparability, we have scaled $\varrho = \rho/16$. As can be seen from this figure, both estimators behave similarly, with Z_ρ interpolating the jumps of the packing number. To the extent that smoother estimators are preferable in many contexts (see the literature on model selection versus shrinkage), it might be that Z_ρ is better behaved. A formalization of this heuristic argument, and a full development of the asymptotic theory of packing numbers, is beyond the scope of the present paper. One advantage of considering Z_ρ, which motivates our focus on this estimator rather than, for instance, the packing number, is that it allows for an easier development of asymptotic theory and of corresponding inference procedures, which are the main object of the present paper.

[10] I thank an anonymous referee for the suggestions discussed in this subsection.

The reader might further wonder, rightly again, whether the sequence of experiments we chose in Assumption 2.3 is peculiar, and whether another sequence might give different answers. The problem of estimating $Z(g)$ might be made more difficult not only by increasing the variance of the regression residuals, but also by letting the roots of g move closer to each other. Formally, we might consider $Y_{i,n} = g_n(X_i) + \epsilon_i$ where (X_i, ϵ_i) are i.i.d. and $g_n(x) = g_0(x \cdot r_n)$ for a diverging sequence r_n. Such a sequence of experiments, however, effectively reduces to the setting of standard asymptotics once we substitute the bandwidth ρ by ρ/r_n, and account for the fact that effective sample size grows only at rate n/r_n. This implies, in particular, that the superconsistency result of Theorem 2.1 also applies to this alternative sequence of experiments, which makes it unsuitable for inference.

3. EXTENSIONS AND APPLICATIONS

In this section, several extensions and applications of the results of Section 2 are presented. Sections 3.1–3.3 discuss, respectively, inference on Z if g is identified by more general moment conditions, inference on Z if the domain and range of g are multidimensional and inference on the number of stable and unstable roots. Sections 3.4 and 3.5 discuss identification and inference for the two applications mentioned in the introduction: static games of incomplete information and stochastic difference equations.

3.1. Conditioning on covariates

In the previous section, inference on $Z(g)$ was discussed for functions g identified by the moment condition

$$g(x) = \text{argmin}_y E_{Y|X}[m(Y - y)|X = x].$$

This subsection generalizes to functions g identified by

$$g(x, w_1) = \text{argmin}_y E_{W_2}\big[E_{Y|X,W}[m(Y - y)|X = x, W_1 = w_1, W_2]\big], \tag{3.1}$$

where the parameter of interest now is $Z(g(\cdot, w_1))$, the number of roots of g in x given w_1. The conditional moment restriction (3.1) can be rationalized by a structural model of the form $Y = h(X, W_1, \epsilon)$, where $\epsilon \perp (X, W_1)|W_2$ and g is defined by

$$g(x, w_1) := \text{argmin}_y E_\epsilon[m(h(x, w_1, \epsilon) - y)].$$

We assume that the joint density of X, W is bounded away from zero on the set $supp(X, W_1) \times supp(W_2)$, where *supp* denotes the compact support of either random vector.

The vector W_2 serves as a vector of control variables. The conditional independence assumption $\epsilon \perp (X, W_1)|W_2$ is also known as 'selection on observables'. The function g is equal to the average structural function if $m(\delta) = \delta^2$, and equal to a quantile structural function if $m_q(\delta) = \delta(q - \mathbf{1}(\delta < 0))$. The average structural function will be of importance in the context of games of incomplete information, as discussed in Section 3.4; quantile structural functions will be used to characterize stochastic difference equations in Section 3.5. When games of incomplete information are discussed in Section 3.4, $W = W_1$ will correspond to the component of public information, which is not excluded from either player's response function.

The inference procedure proposed in the previous section is based upon two steps. First, the function g and its derivative are estimated using local linear m-regression. In the second step, the estimator $(\widehat{g}, \widehat{g'})$ is plugged into the functional $Z_\rho(\cdot,\cdot)$, which is a smooth approximation of the functional $Z(\cdot)$. We can generalize this approach by maintaining the same second step while using more general first-stage estimators $(\widehat{g}, \widehat{g'})$. Equation (3.1) suggests estimating g by a non-parametric sample analogue, replacing the conditional expectation with a local linear kernel estimator of it, and the expectation over W_2 with a sample average. Formally, let $(\widehat{g}(x, w_1), \widehat{g'}(x, w_1)) = \operatorname{argmin}_{a,b} M(a, b, x, w_1)$, where

$$M(a,b,x,w_1) = \frac{1}{n}\sum_j \frac{\sum_i K_\tau(X_i - x, W_{1i} - w_1, W_{2i} - W_{2j}) m(Y_i - a - b(X_i - x))}{\sum_i K_\tau(X_i - x, W_{1i} - w_1, W_{2i} - W_{2j})}. \quad (3.2)$$

An asymptotic normality result can be shown in this context, which generalizes Theorem 2.2. In light of the proof of Theorem 2.2, the crucial step is to obtain a sequence of experiments such that $\widehat{g}$ converges uniformly to g while $\widehat{g'}$ has a non-degenerate limiting distribution. If we obtain an approximation of $\widehat{g'}$ equivalent to the approximation in Assumption 2.2, all further steps of the proof apply immediately. This can be done, using the results of Newey (1994), for the following sequence of experiments.

ASSUMPTION 3.1. *Experiments are indexed by n, and for the nth experiment we observe* $(Y_{i,n}, X_{i,n}, W_{i,n})$ *for* $i = 1, \ldots, n$. *The observations* $(X_{i,n}, Y_{i,n}, W_{i,n})$ *are i.i.d. given n, and*

$$(X_{i,n}, W_{i,n}) \overset{\text{i.i.d.}}{\sim} f_{x,w}(\cdot) \quad (3.3)$$

$$\gamma_{i,n} | (X_{i,n}, W_{i,n}) \sim f_{\gamma|X,W} \quad (3.4)$$

$$Y_{i,n} = g(X_{i,n}, W_{1i,n}) + r_n \gamma_{i,n}. \quad (3.5)$$

THEOREM 3.1 (ASYMPTOTIC NORMALITY, WITH CONTROL VARIABLES). *Under the assumptions of Section 2, but with g identified by (3.1) and the data generated by the model given by Assumption 3.1, if* $r_n = (n\tau^{2+d})^{1/2}$, *where* $d = \dim(X) + \dim(W_1)$, $n\tau^d \to \infty$, $\sqrt{\log(n)}\rho \to 0$ *and* $\tau/\rho^2 \to 0$, *then there exist* $\mu > 0$ *and* V *such that*

$$\sqrt{\frac{\rho}{\tau}}(\widehat{Z} - \mu - Z) \xrightarrow{d} N(0, V).$$

3.2. Higher-dimensional systems

Thus far, only one-dimensional arguments x and one-dimensional ranges for the function g have been considered, where x is the argument over which Z_ρ integrates. All results of Section 2 are easily extended to a higher-dimensional set-up. In particular, assume we are interested in the number of roots of a function g from $\mathbb{R}^d$ to $\mathbb{R}^d$. Generalizing (2.4), we can define $\widehat{Z}$ as

$$\widehat{Z} := \int L_\rho(\widehat{g}) |\det \widehat{g'}|, \quad (3.6)$$

where $(\widehat{g}(\cdot), \widehat{g'}(\cdot))$ are again estimated by local linear m regression, L_ρ is a kernel with support $[-\rho, \rho]^d$, and the integral is taken over the set $\mathscr{X} \subset \mathbb{R}^d$ in the support of g. As in the

one-dimensional case, superconsistency follows from uniform convergence of $(\widehat{g}, \widehat{g'})$. The following theorem, generalizing Theorem 2.2, holds for arbitrary d.

THEOREM 3.2 (ASYMPTOTIC NORMALITY, MULTIDIMENSIONAL SYSTEMS). *Under the assumptions of Section 2, but with* $g : \mathbb{R}^d \to \mathbb{R}^d$ *and* $\widehat{Z}$ *defined by (3.6), if* $r_n = (n\tau^{2+d})^{1/2}$, $n\tau^d \to \infty$, $\sqrt{\log(n)}\rho \to 0$ *and* $\tau/\rho^{d+1} \to 0$, *then there exist* $\mu > 0$, V *such that*

$$\left(\frac{\rho}{\tau}\right)^{d/2}(\widehat{Z} - \mu - Z) \xrightarrow{d} N(0, V).$$

3.3. Stable and unstable roots

Instead of testing for the total number of roots, one might be interested in the number of stable and unstable roots, Z^{s} and Z^{u}. Stable roots are those where g' is negative, and unstable roots are those where g' is positive:

$$\begin{aligned} Z^{\mathrm{s}}(g) &:= |\{x \in \mathscr{X} : g(x) = 0 \text{ and } g'(x) < 0\}| \\ Z^{\mathrm{u}}(g) &:= |\{x \in \mathscr{X} : g(x) = 0 \text{ and } g'(x) > 0\}|. \end{aligned} \tag{3.7}$$

In the multidimensional case, we could more generally consider roots with a given number of positive and negative eigenvalues of g'. We can define smooth approximations of the parameters Z^{s} and Z^{u} as follows:

$$\begin{aligned} Z^{\mathrm{s}}_\rho(g(\cdot), g'(\cdot)) &:= \int_{\mathscr{X}} L_\rho(g(x))|g'(x)|\mathbf{1}(g'(x) < 0)\,dx \\ Z^{\mathrm{u}}_\rho(g(\cdot), g'(\cdot)) &:= \int_{\mathscr{X}} L_\rho(g(x))|g'(x)|\mathbf{1}(g'(x) > 0)\,du. \end{aligned} \tag{3.8}$$

Again, all arguments of Section 2 go through essentially unchanged for these parameters. In particular, Theorem 2.2 applies literally, replacing Z with Z^{s} or Z^{u}.

More generally, functionals that are smooth approximations of the number of roots with various stability properties can be constructed in the multidimensional case by multiplying the integrand with an indicator function depending on the signs of the eigenvalues of $\widehat{g'}$.

3.4. Static games of incomplete information

This section and Section 3.5 discuss how to apply the inference procedure proposed to test for equilibrium multiplicity in economic models. The discussion in this subsection builds on Bajari et al. (2010).

Consider the following static game of incomplete information. Assume there are two players $i = 1, 2$, who both have to choose between one of two actions, $a = 0, 1$. Player i makes her choice based on public information s, as well as private information ϵ_i. The public information s is observed by the econometrician, and ϵ_i is independent of s. It is assumed that ϵ_{-i} does not enter player i's utility.[11] Denote the probability that player i plays strategy $a = 1$ given the

[11] This is an important restriction. It precludes, in particular, application of this set-up to correlated value auctions.

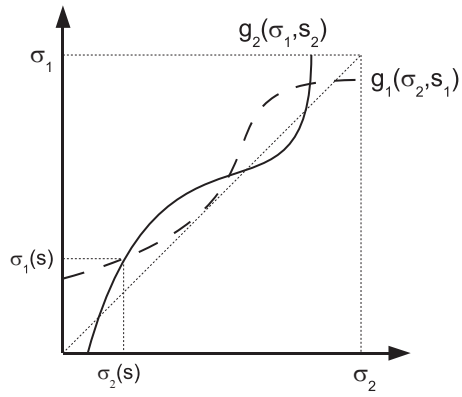

Figure 4. Response functions and Bayesian Nash equilibria.

public information s by $\sigma_i(s)$. Player i's expected utility given her information, and hence her optimal action a_i, depends on s and ϵ_i, as well as player $-i$'s probability of choosing $a = 1$, $\sigma_{-i}(s)$. Let us denote the average best response of player i, integrating over the distribution of ϵ_i, by

$$g_i(\sigma_{-i}, s) = E[a_i|\sigma_{-i}, s]. \tag{3.9}$$

Figure 4 illustrates, by plotting the response functions g_i for a given s. The functions g_i are the (average) best response functions, Bayesian Nash equilibrium requires $g(\sigma_1, s) := g_1(g_2(\sigma_1, s_2), s_1) - \sigma_1 = 0$, and we observe one equilibrium $(\sigma_1(s), \sigma_2(s))$ in the data. In this figure, there are two further equilibria that are not directly observable. In Bayesian Nash equilibrium, the probability of player i choosing $a = 1$, σ_i, equals the average best response of player i, g_i. This implies the two equilibrium conditions

$$\sigma_i(s) = g_i(\sigma_{-i}(s), s),$$

for $i = 1, 2$. In Figure 4, the Bayesian Nash equilibria correspond to the intersections of the graphs of the two g_i. The condition for Bayesian Nash equilibrium in this game can be restated as $g(\sigma_1, s) = 0$, where

$$g(\sigma_1, s) = g_1(g_2(\sigma_1, s), s) - \sigma_1. \tag{3.10}$$

The number of roots of $g(\sigma_1, s)$ in σ_1 is the number of Bayesian Nash equilibria in this game, given s.

We now discuss identification and inference on the number of Bayesian Nash equilibria of this game, given the public information s. Assume we observe an i.i.d. sample of $(a_{1,j}, a_{2,j}, s_j)$, the players' realized actions and the public information of the game, where $a_{i,j} \in \{0, 1\}$ for $i = 1, 2$ and $s \in \mathbb{R}^k$. In this subsection, i indexes players and j indexes observations. Rational expectation beliefs of player $-i$ about the expected action of player i are given by $\sigma_i(s) = E[a_i|s]$. The following two-stage estimation procedure is a non-parametric variant of the

procedure proposed by Bajari et al. (2010). We can get an estimate of the beliefs, $\widehat{\sigma}_i(s) = \widehat{E}[a_i|s]$, by local linear mean regression.

$$(\widehat{\sigma}_i(s), \widehat{\sigma}_i'(s)) = \text{argmin}_{b,c} \sum_j K_\tau(s_j - s)(a_{i,j} - b - c(s_j - s))^2. \tag{3.11}$$

Average best responses of players are given by $g_i(\sigma_{-i}, s) = E[a_i|\sigma_{-i}, s]$. Without further restrictions, g_i is not identified, because by definition σ is functionally dependent on s. If, however, exclusion restrictions of the form

$$g_i(\sigma_{-i}, s) = g_i(\sigma_{-i}, s_i) \tag{3.12}$$

are imposed, g_i can be identified. In particular, assume that exclusion restriction (3.12) holds, with $\dim(s_i) = \dim(s) - 1 = k - 1$. There is one excluded component of s for each player, the remaining $k - 2$ components are not excluded from either response function g_i. Assume furthermore that $\sigma_i(s)$ has full support $[0, 1]$ given s_{-i}, for $i = 1, 2$. Under these assumptions, we can estimate the best response functions, $\widehat{g}_i(\bar{\sigma}_{-i}, s_i) = \widehat{E}[a_i|\widehat{\sigma}_{-i} = \bar{\sigma}_{-i}, s_i]$, again using local linear mean regression:

$$\begin{aligned}(\widehat{g}_i(\bar{\sigma}_{-i}, s_i), \widehat{g'}_i(\bar{\sigma}_{-i}, s_i)) = \text{argmin}_{b,c} \sum_j K_\tau(\widehat{\sigma}_{-i,j} - \bar{\sigma}_{-i}, s_{i,j} - s_i) \\ \times \big(a_{i,j} - b - c(\widehat{\sigma}_{-i,j} - \bar{\sigma}_{-i}, s_{i,j} - s_i)\big)^2.\end{aligned} \tag{3.13}$$

Note that no functional form restrictions are needed for identification of the choice functions g_i. This stands in contrast to Bajari et al. (2010), who need to impose such restrictions in order to be able to identify the underlying preferences. Recall that the condition for Bayesian Nash equilibrium in this game is given by $g(\bar{\sigma}_1, s) = g_1(g_2(\bar{\sigma}_1, s_2), s_1) - \bar{\sigma}_1 = 0$. Inserting $\widehat{g}_2$ into $\widehat{g}_1$, both estimated by (3.13), yields an estimator of g, which can be written as

$$\widehat{g}(\bar{\sigma}_1, s) = \widehat{E}\big[a_1|\widehat{\sigma}_2 = \widehat{E}[a_2|\widehat{\sigma}_1 = \bar{\sigma}_1, s_2], s_1\big] - \bar{\sigma}_1. \tag{3.14}$$

Based on this estimator, we can perform inference on the number of Bayesian Nash equilibria given s, $Z(g(\cdot, s))$. In particular, let

$$\widehat{Z} = Z_\rho(\widehat{g}(\cdot, s), \widehat{g'^1}(\cdot, s)), \tag{3.15}$$

where $\widehat{g}(\cdot, s)$ is given by (3.14). The term $\widehat{g'^1}(\cdot, s)$ refers to the estimated derivative of g w.r.t. $\bar{\sigma}_1$, and similarly for $\widehat{g'^1_1}$ and $\widehat{g'^1_2}$, so that

$$\widehat{g'^1}(\bar{\sigma}_1, s) = \widehat{g'^1_1}(\widehat{g_2}(\bar{\sigma}_1, s_2), s_1) \cdot \widehat{g'^1_2}(\bar{\sigma}_1, s_1). \tag{3.16}$$

Inference on $Z(g(\cdot, s))$ can now proceed as before, if an asymptotic normality result similar to Theorem 2.2 can be shown. In the proof of Theorem 2.2, three properties of $(\widehat{g}(\cdot), \widehat{g'}(\cdot))$ needed to be proven for the statement of the theorem to follow. First, under the given sequence of experiments, $\widehat{g}(\cdot)$ converges uniformly in probability to a degenerate limit. Second, $\widehat{g'}(\cdot)$ converges in distribution to a non-degenerate limit. Third, $\widehat{g'}(x_1)$ and $\widehat{g'}(x_2)$ are asymptotically independent for $|x_1 - x_2| > \text{const} \cdot \tau$. These properties can be shown for $r_n \cdot ((\widehat{g}(\cdot, s), \widehat{g'^1}(\cdot, s))$ in the present case, with $\bar{\sigma}_1$ replacing x, for an appropriate choice of sequence of experiments, where r_n is a scale parameter as before.

The choice of sequence of experiments may seem to be more complicated here than in the baseline case, because the dependent variable a is naturally bounded by [0, 1], so that increasing the residual variance would be inconsistent with the structural model. This is not a problem, however, if we note that the distribution of $\widehat{Z}$, in the baseline model, is invariant to a proportional rescaling of Y, g and ρ. We can therefore define a sequence of experiments that is equivalent to the one defined by (2.7)–(2.9) if we replace (2.9) by

$$Y_{i,n} = \frac{1}{r_n} g(X_{i,n}) + \gamma_{i,n} \tag{2.9'}$$

and ρ by ρ / r_n. Intuitively, shrinking the signal g is equivalent to increasing the noise $r_n \gamma_{i,n}$. Returning to games of incomplete information, consider the following sequence of experiments.

ASSUMPTION 3.2. *For $i = 1, 2$, $g_{i,0}$ is continuously differentiable and monotonic in σ_{-i}, and $g_{i,n}^{-1}$ denotes the inverse of $g_{i,n}$ with respect to the $\sigma_{i,n}$ argument, given s_i. Experiments are indexed by n, and for the nth experiment we observe $(s_j, a_{1,j,n}, a_{2,j,n})$ for $j = 1, \ldots, n$. The observations $(s_j, a_{1,j,n}, a_{2,j,n})$ are i.i.d. given n and*

$$s_{j,n} \sim f_s(\cdot) \tag{3.17}$$

$$a_{i,j,n} | s_{j,n} \sim \text{Bin}(\sigma_{i,n}(s_{j,n})) \tag{3.18}$$

$$\sigma_{i,n}(s) = g_{i,n}(\sigma_{-i,n}(s), s_i) \tag{3.19}$$

$$g_{1,n}(\sigma_2, s_1) = \frac{1}{r_n} g_{1,0}(\sigma_2, s_1) + (1 - \frac{1}{r_n})\sigma_2 \tag{3.20}$$

$$g_{2,n}^{-1}(\sigma_2, s_2) = \frac{1}{r_n} g_{2,0}^{-1}(\sigma_2, s_2) + (1 - \frac{1}{r_n})\sigma_2. \tag{3.21}$$

Equations (3.17)–(3.19) are the same as in the model we have been discussing so far. Equations (3.20) and (3.21) shrink the graphs of the best response functions $g_i(\cdot, s_i)$ towards the $\sigma_1 = \sigma_2$ line (compare Figure 4), parallel to the σ_1 axis. Denote $\sigma_{2,n} = g_{2,n}(\sigma_1, s_2)$. We obtain

$$\begin{aligned} g_n(\sigma_1, s) &= g_{1,n}(g_{2,n}(\sigma_1, s_2), s_1) - \sigma_1 = g_{1,n}(\sigma_{2,n}, s_1) - g_{2,n}^{-1}(\sigma_{2,n}, s_2) \\ &= \frac{1}{r_n}(g_{1,0}(\sigma_{2,n}, s_1) - g_{2,0}^{-1}(\sigma_{2,n}, s_2)). \end{aligned}$$

By (3.21), if $r_n \to \infty$, then $\sigma_{2,n} \to \sigma_1$, and hence

$$r_n g_n(\sigma_1, s) \to g_{1,0}(\sigma_1, s_1) - g_{2,0}^{-1}(\sigma_1, s_2). \tag{3.22}$$

Using this sequence of experiments, we can now state an asymptotic normality result, similar to Theorem 2.2, for static games of incomplete information. The statement of the theorem differs in two respects from the baseline case. First, ρ is replaced by $r_n \rho$ in all expressions. Because this sequence of experiments shrinks g rather than expanding the error, the bandwidth ρ must also shrink correspondingly. Second, the rate of growth of r_n is smaller. Because all regressions are controlling for s_1 or s_2, rates of convergence are slower. In particular, $r_n \cdot \widehat{g_i'^1}$ converges to a

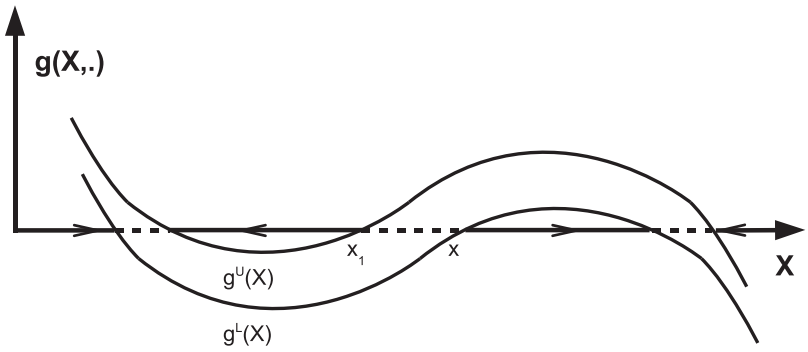

Figure 5. Qualitative dynamics of stochastic difference equations.

non-degenerate limit iff $r_n = O((n\tau^{2+k})^{1/2})$, where k is the dimensionality of the support of the response functions g_i, $k = \dim(s)$.

THEOREM 3.3 (ASYMPTOTIC NORMALITY, STATIC GAMES OF INCOMPLETE INFORMATION). *Under the sequence of experiments defined by Assumption 3.2, if* $R = o_p((\widehat{g}, \widehat{g'}) - (g, g'))$ *uniformly in the Bahadur expansions as* $n \to \infty$, *and if* $r_n = (n\tau^{2+k})^{1/2}$, $n\tau \to \infty$, $r_n\sqrt{\log(n)}\rho \to 0$ *and* $\tau/(r_n\rho)^2 \to 0$, *then there exist* $\mu > 0$ *and* V *such that*

$$\sqrt{\frac{r_n\rho}{\tau}}(\widehat{Z} - \mu - Z) \xrightarrow{d} N(0, V).$$

3.5. Stochastic difference equations

In this subsection, we discuss the identification and interpretation of the number of roots of g for stochastic difference equations of the form

$$\Delta X_{i,t+1} = X_{i,t+1} - X_{i,t} = g(X_{i,t}, \epsilon_{i,t}). \tag{3.23}$$

Interest in such difference equations is motivated by the study of neighbourhood composition dynamics in Card et al. (2008). This discussion will form the basis of the empirical application in Section 4. The results of this subsection suggest that, if the stochastic difference equation (3.23) had multiple equilibria, then we should expect to find multiple roots in cross-sectional quantile regressions of ΔX on X. The notion of multiple equilibria here has to be generalized to the notion of multiple equilibrium regions.

The intuition for this claim is as follows. Holding ϵ constant, the number of roots of g in X is the number of equilibria of the difference equation (3.23). If ϵ is stochastic, then the number of roots can still serve to characterize qualitative dynamics in terms of equilibrium regions. This is shown in Figure 5, which illustrates the characterization of dynamics derived in this section. In this figure, g^{U} and g^{L} are upper and lower envelopes of g for a sequence of realizations of ϵ. There are ranges of X in which the sign of ΔX does not depend on ϵ. This implies that in these ranges X moves towards the equilibrium regions, which are the regions in which the roots of $g(\cdot, \epsilon)$ lie. Equilibrium regions correspond to the dashed segments of the X-axis, the basin of

attraction of the lower equilibrium region is given by $(-\infty, x_1]$ and the basin of attraction of the upper equilibrium region is $[x_2, \infty)$.

How is the joint distribution of (X_t, X_{t+1}) related to the transition function g? Unobserved heterogeneity, which is positively related over time, leads to an upward bias in quantile regression slopes relative to the corresponding structural slopes. To show this, denote the qth conditional quantile of ΔX given X by $Q_{\Delta X|X}(q|X)$, the conditional cumulative distribution function at Q by $F_{\Delta X|X}(Q|X)$, and the conditional probability density by $f_{\Delta X|X}(Q|X)$. The following lemma shows that quantile regressions of ΔX on X yield biased slopes relative to the structural slope $(\partial/\partial X)g$, if X is not exogenous. The second term in (3.24) reflects the bias due to statistical dependence between X and ϵ.

LEMMA 3.1 (BIAS IN QUANTILE REGRESSION SLOPES). *If $\Delta X = g(X, \epsilon)$, and if Q and F are differentiable with respect to the conditioning argument X, then*

$$\frac{\partial}{\partial X} Q_{\Delta X|X}(\tau|X) = E[\frac{\partial}{\partial X} g(X, \epsilon)|\Delta X = Q, X] \\ - \frac{1}{f_{\Delta X|X}(Q|X)} \cdot \frac{\partial}{\partial X} \mathbb{P}(g(X', \epsilon) \leq Q|X)|_{X'=X}. \tag{3.24}$$

The following assumption of first-order stochastic dominance states that there is no negative dependence between current $g(x', \epsilon)$, evaluated at fixed x', and current X.

ASSUMPTION 3.3 (FIRST-ORDER STOCHASTIC DOMINANCE). *$\mathbb{P}(g(x', \epsilon) \leq Q|X)$ is non-increasing as a function of X, holding x' constant.*

Violation of this assumption would require some underlying cyclical dynamics, in continuous time, with a frequency close enough to half the frequency of observation, or more generally with a ratio of frequencies that is an odd number divided by two. It seems safe to discard this possibility in most applications. This assumption might not hold, for instance, if outcomes were influenced by seasonal factors and observations were semi-annual.

We can now formally state the claim that, if there are unstable equilibria structurally, then quantile regressions should exhibit multiple roots.

PROPOSITION 3.1 (UNSTABLE EQUILIBRIA IN DYNAMICS AND QUANTILE REGRESSIONS). *Assume that $\Delta X = g(X, \epsilon)$ and that $g(\inf \mathscr{X}, \epsilon) > 0$, and $g(\sup \mathscr{X}, \epsilon) < 0$ for all ϵ. If Assumption 3.3 holds and $Q_{\Delta X|X}(q|X)$ has only one root X for all q, then the conditional average structural functions $E[g(x', \epsilon)|g(X, \epsilon) = 0, X]$, as functions of x', are stable at the roots m:*

$$E[\frac{\partial}{\partial X} g(X, \epsilon)|\Delta X = 0, X] \leq 0$$

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

This proposition assumes global stability of g (i.e. X does not diverge to infinity). Under such global stability, if there is only one root of g, then this root is stable. According to this proposition, if quantile regressions only have one stable root, then the same is true for the conditional average structural functions. This is not conclusive, but it is suggestive that $g(\cdot, \epsilon)$ themselves have only one root.

Let us now turn to the implications of the number of roots of g for the qualitative dynamics of the stochastic difference equation (3.23). Let $\tilde{g}(x, \epsilon) := g(x, \epsilon) + x$. If g describes a structural

relationship, the counterfactual time path under 'manipulated' initial condition $X_{i,0} = x'$ is given by

$$\begin{aligned} X_{i,1} &= \tilde{g}(x', \epsilon_{i,0}) \\ X_{i,2} &= \tilde{g}(X_{i,1}, \epsilon_{i,1}) \\ &\vdots \\ X_{i,t} &= \tilde{g}(X_{i,t-1}, \epsilon_{i,t-1}). \end{aligned} \tag{3.25}$$

Given the initial condition $X_{i,1}$ and shocks $\epsilon_{i,1}, \ldots, \epsilon_{i,t}$, (3.23) describes a time inhomogeneous deterministic difference equation. The following argument makes statements about the qualitative behaviour of this difference equation based on properties of the function g, in particular based on the number of roots in x of $g(x, \epsilon)$ for given unobservables $\epsilon_{i,1}, \ldots, \epsilon_{i,t}$. Consider Figure 5, which shows g^{U} and g^{L} defined by

$$g^{\mathrm{U}}_{i,t}(x) = \max_{0 \le s < t} g(x, \epsilon_{i,s}) \tag{3.26}$$

$$g^{\mathrm{L}}_{i,t}(x) = \min_{0 \le s < t} g(x, \epsilon_{i,s}). \tag{3.27}$$

The functions $g^{\mathrm{U}}_{i,t}$ and $g^{\mathrm{L}}_{i,t}$ are the upper and lower envelopes of the family of functions $g(x, \epsilon_{i,s})$ for $s = 1, \ldots, t$. The direction of movement of X over time does not depend on s in the ranges where $g^{\mathrm{U}}_{i,t} < 0$ or $g^{\mathrm{L}}_{i,t} > 0$ (which is where the horizontal axis is drawn solid in Figure 5), because the sign of $g(x, \epsilon_{i,s})$ does not depend on s in these ranges. In other words, suppose we start off with an initial value below x_1 in the picture. If that is the case, $X_{i,s}$ will converge monotonically toward the left-hand dashed range and then remain within that range for all $s \le t$. Similarly, for $X_{i,0}$ in the upper 'basin of attraction' beyond x_2, $X_{i,s}$ will converge to the upper 'equilibrium range' given by the right-hand dashed range. Hence, small changes of initial conditions (from x_1 to x_2) can have large and persistent effects on X in this case, in contrast to the case where $g(\cdot, \epsilon)$ only has one stable root for all ϵ. These arguments are summarized in the following proposition.

PROPOSITION 3.2 (CHARACTERIZING DYNAMICS OF STOCHASTIC DIFFERENCE EQUATIONS). *Assume that $g^{\mathrm{U}}_{i,t}$ and $g^{\mathrm{L}}_{i,t}$, defined by* (3.26) *and* (3.27)*, are smooth and generic, positive for sufficiently small x and negative for sufficiently large x, and have the same number z of roots, $x^{\mathrm{U}}_1 < \ldots < x^{\mathrm{U}}_z$ and $x^{\mathrm{L}}_1 < \ldots < x^{\mathrm{L}}_z$, and let $x^{\mathrm{L}}_0 = -\infty$, $x^{\mathrm{U}}_{z+1} = \infty$. Define the following mutually disjoint ranges:*

$$\begin{aligned} N_c &= [x^{\mathrm{U}}_c, x^{\mathrm{U}}_{c+1}] \text{ for } c = 1, 3, \ldots, z; \\ P_c &= [x^{\mathrm{L}}_c, x^{\mathrm{L}}_{c+1}] \text{ for } c = 0, 2, \ldots, z-1; \\ S_c &= [x^{\mathrm{L}}_c, x^{\mathrm{U}}_c] \text{ for } c = 1, 3, \ldots, z; \\ U_c &= [x^{\mathrm{U}}_c, x^{\mathrm{L}}_c] \text{ for } c = 2, 4, \ldots, z-1. \end{aligned}$$

Then, all $g(x, \epsilon_{i,s})$ are negative on the N_c, and positive on the P_c. Furthermore, all $g(x, \epsilon_{i,s})$ are negative in a neighbourhood to the right of the maximum of the S_c and positive to the left of the minimum, and the reverse holds for the U_c. Therefore, if $X_{i,0} \in N_c$ and $S_c \neq \varnothing$, then $X_{i,s}$ will converge monotonically toward S_c and then remain within S_c. If $X_{i,0} \in P_c$ and $S_{c+1} \neq \varnothing$, then $X_{i,s}$ will converge monotonically toward S_{c+1} and then remain within S_{c+1}.

Assuming non-emptiness of these ranges, the interval $P_{c-1} \cup S_c \cup N_c$ is a basin of attraction for S_c (i.e. X in this interval converges monotonically to S_c and then remains there). The main difference relative to the deterministic, time homogenous case is the blurring of the stable equilibrium to a stable set S_c.

We did not make any assumptions on the joint distribution of the unobserved factors $\epsilon_{i,1}, \ldots, \epsilon_{i,t}$. The whole argument of the preceding theorem is conditional on these factors. However, the predictions of the theorem will be sharper (given g) if serial dependence of unobserved factors is stronger, increasing the number of units i to which the assertion is applicable and reducing the size of the intervals S_c and U_c, because $g^{\mathrm{U}}_{i,t} - g^{\mathrm{L}}_{i,t}$ is going to be smaller on average.

In summary, Proposition 3.1 implies that, if we do not find multiple roots in quantile regressions, then the conditional average structural functions $E[g(x', \epsilon)|g(X, \epsilon) = 0, X]$ do not have multiple roots. Proposition 3.2 implies that, if upper and lower envelopes of $g(\cdot, \epsilon_{i,s})$ do not have multiple roots, then the dynamics of the system are stable and initial conditions do not matter in the long run.

4. APPLICATION TO THE DYNAMICS OF NEIGHBOURHOOD COMPOSITION

This section analyses the dynamics of minority share in a neighbourhood, applying the methods developed in the last two sections to the data used for analysis of neighbourhood composition dynamics by Card et al. (2008). They study whether preferences over neighbourhood composition lead to a 'white flight', once the minority share in a neighbourhood exceeds a certain level. They argue that such 'tipping' behaviour implies discontinuities in the change of neighbourhood composition over time as a function of initial composition, and they test for the presence of such discontinuities in cross-sectional regressions over different neighbourhoods in a given city. This argument is based on the theoretical models of Becker and Murphy (2000), which do not allow for individual heterogeneity and consider infinite time horizons. The present paper argues that, if we allow for heterogeneity and finite time, and if tipping does take place, then we should expect multiple roots rather than discontinuities. Kasy (2015) discusses a search-matching model of the housing market with social externalites, which has this implication.

Card et al. (2008) provided full access to their datasets, which allows us to use identical samples and variable definitions as in their work. The dataset is an extract from the Neighbourhood Change Database (NCDB), which aggregates US census variables to the level of census tracts. Tract definitions are changing between census waves but the NCDB matches observations from the same geographic area over time, thus allowing observation of the development over several decades of the universe of US neighbourhoods. In the dataset used by Card et al. (2008), all rural tracts are dropped, as well as all tracts with population below 200 and tracts that grew by more than five standard deviations above the metropolitan statistical area (MSA) mean. The definition of MSA used is the MSAPMA from the NCDB, which is equal to a 'primary metropolitan statistical area' if the tract lies in one of those, and equal to the MSA it lies otherwise. For further details on sample selection and variable definition, see Card et al. (2008).

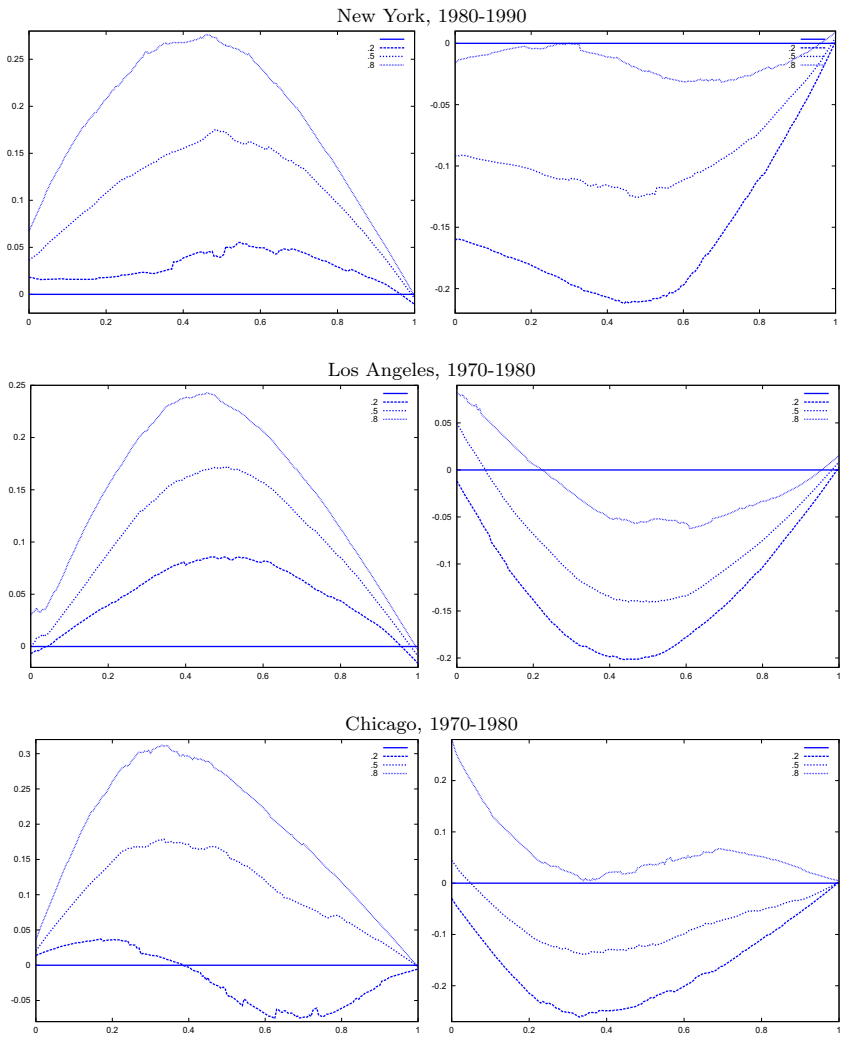

Figure 6. Quantile regressions of the changes in minority share and white population.

The graphs and tables to be discussed are constructed as follows. For each of the MSAs and each of the decades separately, we run local linear quantile regressions of the change in minority share of a neighbourhood (tract) on minority share at the beginning of the decade. This is done for the quantiles 0.2, 0.5 and 0.8, with a bandwidth τ of $n^{-0.2}$, where n is the sample size.[12] Figure 6 shows local linear quantile regressions of the change in minority share (left

[12] The implementation of local linear quantile regression uses code downloaded from http://www.econ.uiuc.edu/~roger/research/rq/rq.html.

Table 1. 0.95 confidence sets for $Z(g)$ by decade and quantile, change in minority share.

	1970s			1980s			1990s		
MSA	$q = 0.2$	$q = 0.5$	$q = 0.8$	$q = 0.2$	$q = 0.5$	$q = 0.8$	$q = 0.2$	$q = 0.5$	$q = 0.8$
New York, NY PMSA	[0,1]	[0,1]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]
Los Angeles-Long Beach, CA PMSA	[1,1]	[1,1]	[0,1]	[0,1]	[0,1]	[0,1]	[1,1]	[1,1]	[0,0]
Chicago, IL PMSA	[0,1]	[0,1]	[0,1]	[2,2]	[0,1]	[0,1]	[1,1]	[0,1]	[0,0]
Dallas, TX PMSA	[1,2]	[1,1]	[0,0]	[0,1]	[0,0]	[0,0]	[0,1]	[0,1]	[0,0]
Philadelphia, PA-NJ PMSA	[1,2]	[0,1]	[0,1]	[1,1]	[0,1]	[0,1]	[1,1]	[0,1]	[0,0]
Houston, TX PMSA	[1,1]	[0,0]	[0,0]	[1,2]	[0,1]	[0,0]	[0,1]	[0,0]	[0,0]
Miami, FL PMSA	[0,1]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]
Washington, DC-MD-VA-WV PMSA	[0,1]	[0,0]	[0,0]	[1,1]	[0,1]	[0,0]	[1,1]	[0,1]	[0,0]
Atlanta, GA MSA	[1,1]	[1,1]	[0,0]	[2,3]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]
Boston, MA-NH PMSA	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,0]	[1,1]	[0,0]	[0,1]
Detroit, MI PMSA	[1,2]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,0]
Phoenix-Mesa, AZ MSA	[1,1]	[0,0]	[0,0]	[1,1]	[0,1]	[0,0]	[1,1]	[0,1]	[0,0]
San Francisco, CA PMSA	[1,1]	[0,1]	[0,1]	[0,0]	[0,1]	[0,0]	[1,1]	[0,0]	[0,0]

Note: The table shows confidence intervals in the integers for $Z(g)$ for the 12 largest MSAs of the United States, ordered by population, where g is estimated by quantile regression of the change in minority share over a decade on the initial minority share for the quantiles 0.2, 0.5 and 0.8. Regression bandwidth τ is $n^{-0.2}$, and σ is chosen as 0.04. Confidence sets are based on t-statistics using bootstrapped bias and standard errors.

Table 2. 0.95 confidence sets for $Z(g)$ by decade and quantile, change in white population.

	1970s			1980s			1990s		
MSA	$q = 0.2$	$q = 0.5$	$q = 0.8$	$q = 0.2$	$q = 0.5$	$q = 0.8$	$q = 0.2$	$q = 0.5$	$q = 0.8$
New York, NY PMSA	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]
Los Angeles-Long Beach, CA PMSA	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]
Chicago, IL PMSA	[0,1]	[0,1]	[0,1]	[0,0]	[0,1]	[1,1]	[0,1]	[0,1]	[0,1]
Dallas, TX PMSA	[0,1]	[0,1]	[0,1]	[0,0]	[1,1]	[0,2]	[0,1]	[1,1]	[0,1]
Philadelphia, PA-NJ PMSA	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[1,1]
Houston, TX PMSA	[0,1]	[0,1]	[0,1]	[1,1]	[1,1]	[1,1]	[0,1]	[0,1]	[0,1]
Miami, FL PMSA	[0,1]	[0,1]	[0,1]	[0,0]	[0,0]	[1,1]	[1,1]	[1,1]	[1,1]
Washington, DC-MD-VA-WV PMSA	[0,1]	[0,0]	[0,1]	[0,0]	[1,1]	[0,0]	[0,1]	[0,1]	[0,1]
Atlanta, GA MSA	[0,1]	[1,1]	[0,1]	[1,1]	[1,1]	[1,1]	[1,1]	[1,2]	[0,1]
Boston, MA-NH PMSA	[0,1]	[0,1]	[0,1]	[0,0]	[0,0]	[1,1]	[0,0]	[0,1]	[0,1]
Detroit, MI PMSA	[0,1]	[0,1]	[0,1]	[0,0]	[0,0]	[1,1]	[0,1]	[0,1]	[0,1]
Phoenix-Mesa, AZ MSA	[0,1]	[0,1]	[0,1]	[0,0]	[1,1]	[0,0]	[0,1]	[0,1]	[0,1]
San Francisco, CA PMSA	[0,1]	[0,1]	[0,1]	[0,0]	[0,0]	[0,0]	[0,0]	[1,1]	[0,0]

Note: The table shows confidence intervals in the integers for $Z(g)$ for the 12 largest MSAs of the United States, ordered by population, where g is estimated by quantile regression of the change in the non-Hispanic, white population over a decade, divided by initial total population, on the initial minority share for the quantiles 0.2, 0.5 and 0.8. Regression bandwidth τ is $n^{-0.2}$, and σ is chosen as 0.05 times the maximal change. Confidence sets are based on t-statistics using bootstrapped bias and standard errors.

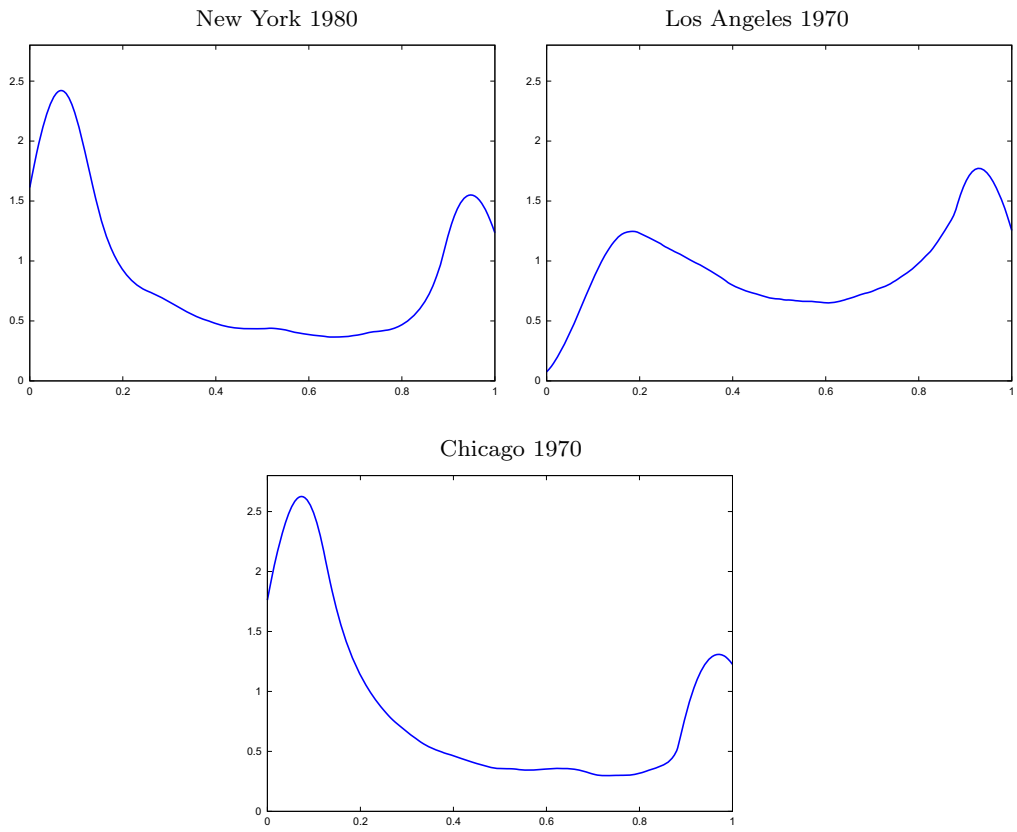

Figure 7. Density of minority share across neighbourhoods.

column) and of the change in white population relative to initial population (right column) on initial minority share for the quantiles 0.2, 0.5 and 0.8. The figures do not show confidence bands. The figure plots these quantile regressions for the three largest MSAs. For each of the regressions, Z_ρ is calculated, where ρ is chosen as 0.04. The integral in the expression for Z_ρ is taken over the interval [0, 1], intersected with the support of initial minority share if the latter is smaller. Note that it is possible to find no (stable) equilibrium for an MSA (i.e. $Z_\rho < 1$), if high initial minority shares do not occur in that MSA and most neighbourhoods experienced growing minority shares. Figure 7 shows kernel density plots of the regressor, the initial minority share across neighbourhoods, which suggest that support problems are not an issue, at least for the largest MSAs. For each Z_ρ, bootstrap standard errors and bias are calculated, as well as the corresponding t-test statistics for the null hypothesis $Z_\rho = 0, 1, 2, 3, \ldots$, implying an integer-valued confidence set (of level 0.05) for z. By the results of Section 2, these confidence sets have an asymptotic coverage probability of 95%. By the Monte Carlo evidence of Appendix A, they are likely to be conservative (i.e. have a larger coverage probability). If the confidence sets thus obtained are empty, the two neighbouring integers of $\widehat{Z}$ are included in the intervals shown. This makes inference even more conservative. Table 1 shows the resulting confidence sets for the

12 largest MSAs in the United States (by 2009 population), for all quantiles and decades under consideration.[13]

As can be seen from the table, in very few cases there is evidence of Z exceeding 1. In all cases shown, except for the 0.2 quantile for Atlanta in the 1980s, we can reject the null $Z \geq 3$. Similar patterns hold for almost all of the 118 cities in the dataset. Rather than exhibiting multiple equilibria, the data indicate a general rise in minority share that is largest for neighbourhoods with intermediate initial share, but not to the extent of leading to tipping behaviour. Proposition 3.1 suggests that, if we do not find multiple roots in quantile regressions, we can reject multiple equilibria in the underlying structural relationship. I take these results as indicative that tipping is not a widespread phenomenon in US ethnic neighbourhood composition over the decades under consideration. This stands in contrast to the conclusion of Card et al. (2008), who do find evidence of tipping.

The approach used here differs from the main analysis in Card et al. (2008) in a number of ways. Card et al. (2008) (a) use polynomial least-squares regression with a discontinuity. They (b) use a split sample method to test for the presence of a discontinuity, and they (c) regress the change in the non-Hispanic, white population, divided by initial neighbourhood population, on initial minority share. We (a) use local linear quantile regression without a discontinuity, we (b) run the regressions on full samples for each MSA and test for the number of roots, and we (c) regress the change in minority share on initial minority share.

To check whether the differing results are due to variable choice (c) rather than testing procedure, the left column of Figure 6 and Table 1 are replicated using the change in the non-Hispanic, white population relative to initial population as the dependent variable, as did Card et al. (2008). The right column of Figure 6 shows such quantile regressions. These figures correspond to the ones in Card et al. (2008, p. 190), using the same variables but a different regression method and the full samples. Table 2 shows confidence sets for the number of roots of these regressions for the 12 largest MSAs. In comparing Tables 1 and 2, note that there is a correspondence between the lower quantiles of the first (low increase in minority share) and the upper quantiles of the latter (higher increase/lower decrease of white population). The two tables show fairly similar results. Again, no systematic evidence of multiple roots is found.

Some factors might lead to a bias in the estimated number of equilibria, using the methods developed here. First, the test might be sensitive to the chosen range of integration if there are roots near the boundary. If a root lies right on the boundary of the chosen range of integration, it enters Z_ρ as $1/2$ only. Extending the range of integration beyond the unit interval, however, might also lead to an upward bias in the estimated number of roots, if extrapolated regression functions intersect with the horizontal axis. Second, choosing a bandwidth parameter ρ that is too large might bias the estimated number of equilibria downwards, if the function g peaks within the range $[-\rho, \rho]$. Third, there might be roots of g in the unit interval but beyond the support of the data.

5. SUMMARY AND CONCLUSION

This paper proposes an inference procedure for the number of roots of functions non-parametrically identified using conditional moment restrictions, and develops the corresponding

[13] The full set of results for all 115 MSAs in the dataset can be found in the online Appendix.

asymptotic theory. In particular, it is shown that a smoothed plug-in estimator of the number of roots is superconsistent under i.i.d. asymptotics, but asymptotically normal under non-standard asymptotics, and asymptotically efficient relative to a simple plug-in estimator. In Section 3, these results are extended to cover various more general cases, allowing for covariates as controls, higher-dimensional domain and range, and for inference on the number of equilibria with various stability properties. This section also discusses how to apply the results to static games of incomplete information and to stochastic difference equations. In an application of the methods developed here to data on neighbourhood composition dynamics in the United States, no evidence of multiple of equilibria is found.

The inference procedure can also be used to test for bifurcations (i.e. (dis)appearing equilibria as a function of changing exogenous covariates). It is easy to test the hypothesis $Z(g(\cdot, W_1)) = Z(g(\cdot, W_2))$, because the corresponding estimators $\widehat{Z}(g(\cdot, W_i))$ are independent for W_1 and W_2 further apart than twice the bandwidth τ. If there are bifurcations, small exogenous shifts might have a large (discontinuous) effect on the equilibrium attained, if the 'old' equilibrium disappears.

In the dynamic set-up, one might furthermore consider to apply the procedure to detrended data (e.g. by demeaning ΔY). It seems likely that regressions of detrended data have a higher number of roots. The rationale of such an approach could be found in underlying models in which the dynamics of a detrended variable are stationary. This is, in particular, the case in Solow-type growth models, in which GDP or capital stock is stationary after normalizing by a technological growth factor.

Finally, it might also be interesting to extend the results obtained here to cover further cases where g cannot be directly estimated using conditional moment restrictions. The crucial step for such extensions, as illustrated by the various cases discussed in Section 3, is to find a sequence of experiments such that the first-stage estimator $\widehat{g}$ converges in probability to a degenerate limit whereas $\widehat{g'}$ converges in distribution to a non-degenerate limit. Furthermore, $\widehat{g'}(x_1)$ needs to be asymptotically independent of $\widehat{g'}(x_2)$ for all $|x_1 - x_2| > \text{const} \cdot \tau$. There are many potential applications of the results obtained here, where it might be interesting to know whether the underlying dynamics or strategic interactions imply multiple equilibria. Examples include household level poverty traps, intergenerational mobility, efficiency wages, macro models of economic growth (as analysed in the online Appendix), financial market bubbles (herding), market entry and social norms.

ACKNOWLEDGEMENTS

I thank seminar participants at UC Berkeley, UCLA, USC, Brown, NYU, UPenn, LSE, UCL, Sciences Po, TSE, Mannheim and IHS Vienna for their helpful comments and suggestions. I particularly thank Tim Armstrong, David Card, Kiril Datchev, Victor Chernozhukov, Jinyong Hahn, Michael Jansson, Bryan Graham, Susanne Kimm, Patrick Kline, Rosa Matzkin, Enrico Moretti, Denis Nekipelov, James Powell, Alexander Rothenberg, Jesse Rothstein, James Stock and Mark van der Laan for many valuable discussions, and David Card, Alexander Mas and Jesse Rothstein for the access provided to their data. This work was supported by a DOC fellowship from the Austrian Academy of Sciences at the Department of Economics, UC Berkeley.

REFERENCES

Andrews, D. W. and X. Cheng (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica 80*, 2153–211.

Aradillas-Lopez, A. (2010). Semiparametric estimation of a simultaneous game with incomplete information. *Journal of Econometrics 157*, 409–31.

Armstrong, T. B. (2014). Weighted KS statistics for inference on conditional moment inequalities. *Journal of Econometrics 181*, 92–116.

Azariadis, C. and J. Stachurski (2005). Poverty traps. In P. Aghion and S. N. Durlauf (Eds.), *Handbook of Economic Growth*, *Volume 1*, 295–384. Amsterdam: Elsevier.

Bajari, P., H. Hong, J. Krainer and D. Nekipelov (2010). Estimating static models of strategic interactions. *Journal of Business and Economic Statistics 28*, 469–82.

Becker, G. and K. Murphy (2000). *Social Economics: Market Behavior in a Social Environment*. Cambridge, MA: Harvard University Press.

Berry, S. (1992). Estimation of a model of entry in the airline industry. *Econometrica 60*, 889–917.

Bowles, S., S. Durlauf and K. Hoff (2006). *Poverty Traps*. Princeton, NJ: Princeton University Press.

Bresnahan, T. and P. Reiss (1991). Entry and competition in concentrated markets. *Journal of Political Economy 99*, 977–1009.

Card, D., A. Mas and J. Rothstein (2008). Tipping and the dynamics of segregation. *Quarterly Journal of Economics 123*, 177–218.

Choirat, C. and R. Seri (2012). Estimation in discrete parameter models. *Statistical Science 27*, 278–93.

Dasgupta, P. and D. Ray (1986). Inequality as a determinant of malnutrition and unemployment: theory. *Economic Journal 96*(384), 1011–34.

De Paula, A. and X. Tang (2012). Inference of signs of interaction effects in simultaneous games with incomplete information. *Econometrica 80*, 143–72.

Fischer, N., E. Mammen, and J. Marron (1994). Testing for multimodality. *Computational Statistics and Data Analysis 18*, 499–512.

Giné, E., D. Mason and A. Zaitsev (2003). The l^1-norm density estimator process. *Annals of Probability 31*, 719–68.

Hoeffding, W. and H. Robbins (1994). The central limit theorem for dependent random variables. In N. I. Fisher and P. K. Sen (Eds.), *The Collected Works of Wassily Hoeffding*, 205–213. New York, NY: Springer.

Horowitz, J. (2001). The bootstrap. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume *5*, 3159–228. Amsterdam: North-Holland.

Horváth, L. (1991). On L_p-norms of multivariate density estimators. *Annals of Statistics 19*, 1933–49.

Imbens, G. and J. Wooldridge (2007). What's new in econometrics? Weak instruments and many instruments. *NBER Lecture Notes 13, Summer 2007*.

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

Kong, E., O. Linton and Y. Xia (2010). Uniform Bahadur representation for local polynomial estimates of m-regression and its application to the additive model. *Econometric Theory 26*, 1–36.

Lewbel, A. and X. Tang (2011). Identification and estimation of games with incomplete information using excluded regressors. Working paper, Boston College.

Mas-Colell, A., M. Whinston and J. Green (1995). *Microeconomic Theory*. New York, NY: Oxford University Press.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory 10*, 233–53.

Quah, D. (1996). Empirics for economic growth and convergence. *European Economic Review 40*, 1353–75.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*, 557–86.

van derVaart, A. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Young, H. (2008). Social norms. In S. Durlauf and L. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed.). Basingstoke: Palgrave Macmillan.

APPENDIX A: MONTE CARLO EVIDENCE

This section presents simulation results to check the accuracy in finite samples of the asymptotic approximations obtained in Theorem 2.2. In all simulations, X are i.i.d. draws of Uni[0, 1] random variables, and the additive errors γ are either uniformly or normally distributed,

$$\begin{aligned} X_i &\overset{\text{i.i.d.}}{\sim} \text{Uni}[0,1] \\ \gamma_i | X_i &\sim f_{\gamma|X} \\ Y_i &= g^j(X_i) + \gamma_i, \end{aligned} \tag{A.1}$$

where $f_{\gamma|X}$ is an appropriately centred and scaled uniform or normal distribution. Two functions g^j are considered, the first with one root and the second with three roots:

$$g^1(x) = 0.5 - x$$
$$g^2(x) = 0.5 - 5x + 12x^2 - 8x^3.$$

The function g is estimated by median regression, mean regression and 0.9 quantile regression, where the γ in the simulations are shifted appropriately to have median, mean or 0.9 quantile at the respective g. Figures A.1–A.3 and Table A.1 show sequences of four experiments with 400, 800, 1,600 and 3,200 observations. These models are chosen to be comparable to the empirical application discussed in Section 4. The variance of γ in each experiment is chosen to yield the same variance for $\widehat{g'}$, as implied by the asymptotic approximation of the Bahadur expansion, in all experiments for a given g. By the proof of Theorem 2.2, we should therefore obtain similar simulation results across all set-ups. Furthermore, the variance of $\widehat{Z}$ should be constant up to a factor τ/ρ. The parameters of these simulations are chosen to lie in an intermediate range where variation in $\widehat{g'}$ is existent but moderate.

Figure A.1 shows density plots for $\widehat{Z}$ from the sequences of Monte Carlo experiments with uniform errors and g identified by median regression, as described in this appendix; in the online Appendix, similar figures are presented for the other experiments. The upper graph shows the distribution from four experiments with increasing sample size n and correspondingly growing variance of the residual γ, where the true parameter Z equals one. The same holds for the lower graph, except that $Z = 3$. As predicted by Theorem 2.2, biases are positive, and both bias and variance are decreasing in n. Figure A.2 shows the distribution of the naive plug-in estimator $Z(\widehat{g})$, from the same simulations as in Figure A.1. It was shown in Section 2 that this estimator is asymptotically inefficient relative to the smoothed plug-in estimator. This relative inefficiency is reflected in a larger dispersion in the simulations, as can be seen by comparing Figures A.1 and A.2. Figure A.3 shows density plots for $\widehat{Z}$, normalized by its sample mean and standard deviation, from the same simulations as in Figure A.1. It also shows, as a reference, the density of a standard normal. These plots suggest that the sample distribution of $\widehat{Z}$ is somewhat right-skewed relative to a normal distribution.

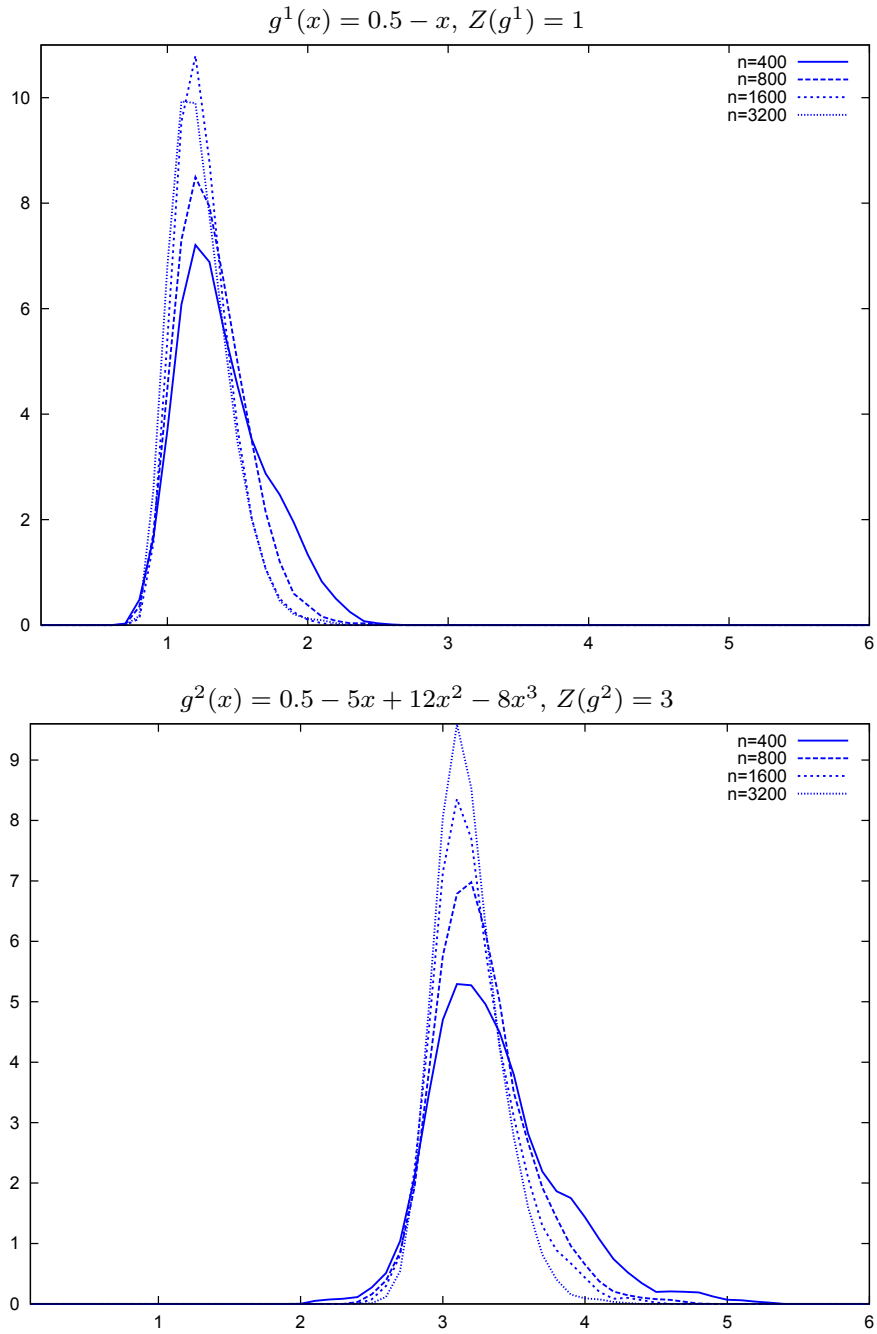

Figure A.1. Density of $\widehat{Z}$ in Monte Carlo experiments.

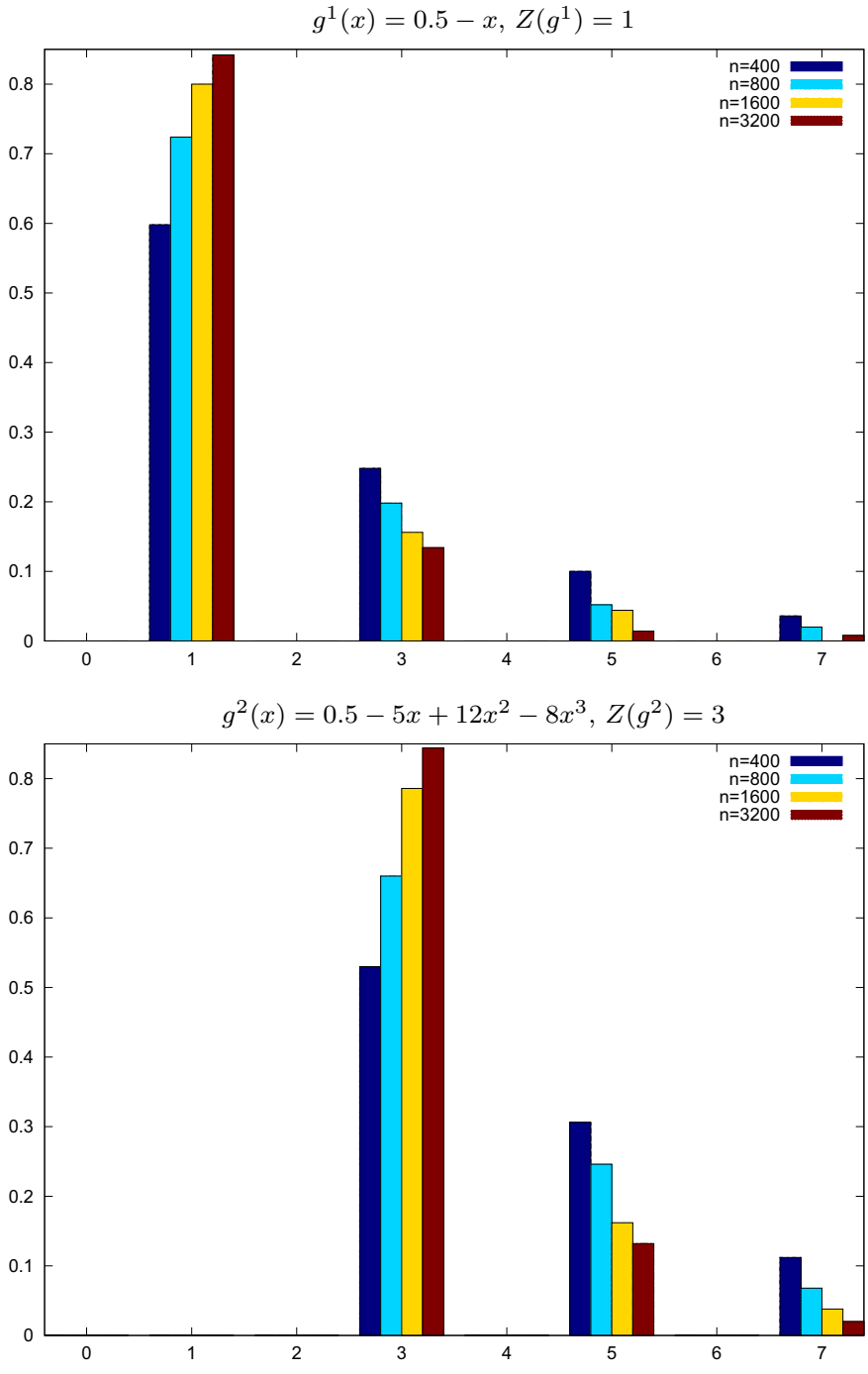

Figure A.2. Distribution of simple plug-in estimator $Z(\widehat{g})$ in Monte Carlo experiments.

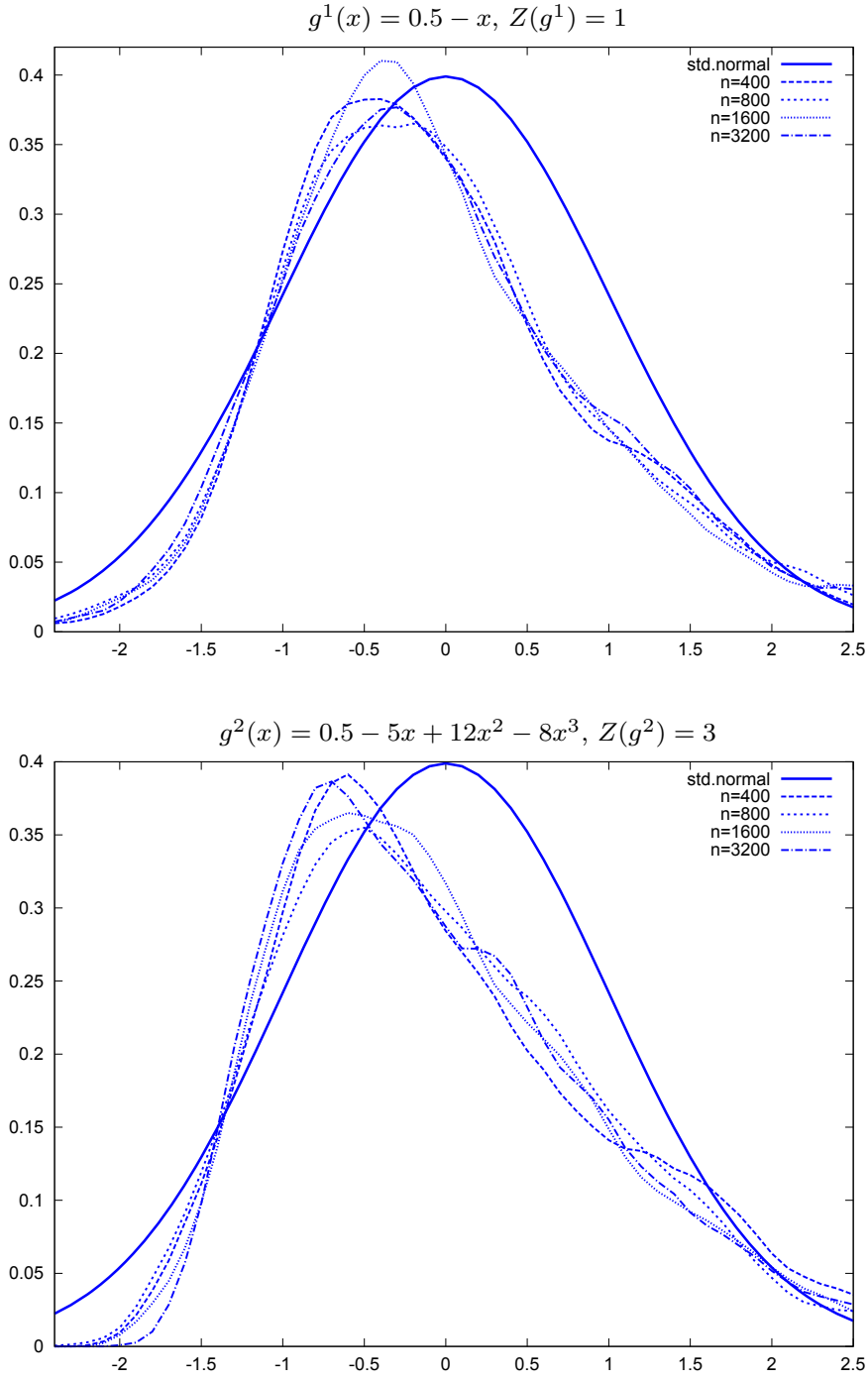

Figure A.3. Density of normalized $\widehat{Z}$ in Monte Carlo experiments.

Table A.1. Monte Carlo rejection probabilities.

n	τ	r	$\widehat{P}(\zeta > z_\alpha)$	$\widehat{P}(\zeta < -z_\alpha)$
400	0.065	0.179	0.05	0.01
800	0.059	0.194	0.03	0.02
1,600	0.055	0.231	0.02	0.01
3,200	0.052	0.290	0.02	0.01
400	0.065	0.268	0.03	0.02
800	0.059	0.292	0.01	0.02
1,600	0.055	0.347	0.01	0.01
3,200	0.052	0.434	0.01	0.02

Note: This table shows the frequency of rejection of the null under a test of asymptotic level 5%, for the sequences of Monte Carlo experiments described in Appendix A. The g are estimated by mean regression, the errors are uniformly distributed, and the first four experiments are generated using g^1 with one root, the next four using g^2 with three roots. The columns show sample size, regression bandwidth, error standard deviation and the rejection probabilities of one-sided tests, respectively.

Table A.1 shows the results of simulations using bootstrapped standard deviations and biases, for mean regression with uniform errors. The results show, for the range of experiments considered, that rejection frequencies are lower than the 0.05 value implied by asymptotic theory. If this pattern generalizes, inference based upon the t-statistic proposed in this paper is conservative in finite samples. In particular, it seems that bootstrapped standard errors are too large.

APPENDIX B: PROOFS

Proof of Proposition 2.1: By continuity of g' as well as genericity of g, we can choose ρ small enough such that $sgn(g'(x))$ is constantly equal to $sgn(g'(x_c)) \neq 0$ in each of the neighbourhoods of the $c = 1, \ldots, z$ roots of g, $\{x_c\}$, defined by $L_\rho(g(x)) \neq 0$. Hence, we can write the integral $\int_{\mathscr{X}} L_\rho(g(x))|g'(x)|dx$ as a sum of integrals over these neighbourhoods, in each of which there is exactly one root. Assume w.l.o.g. that $z = 1$ and $sgn(g')$ is constant in the range of x where $L_\rho(g(x)) \neq 0$. Then, by a change of variables setting $y = g(x)$,

$$\int_{\mathscr{X}} L_\rho(g(x))|g'(x)|dx = \int_{g(\mathscr{X})} L_\rho(y)|g'(g^{-1}(y))|\frac{1}{|g'(g^{-1}(y))|}dy = 1. \qquad \square$$

Proof of Proposition 2.2: We need to find ϵ such that $||g - \tilde{g}|| < \epsilon$ implies $Z(\tilde{g}) = Z(g)$. By genericity of g, each root x_c of g is such that $sgn(g'(x_c)) \neq 0$. By continuous first derivatives, we can then find δ such that $sgn(g'(\cdot))$ is constant in the neighbourhood $NH_c := (x_c - \delta, x_c + \delta)$ of each of the finitely many roots x_c and the NH_c are mutually disjoint. By continuity of g,

$$\epsilon_1 := \inf_{x \notin \bigcup_c NH_c} g(x) > 0 \tag{B.1}$$

and

$$\epsilon_2 := \inf_{x \in \bigcup_c \bar{NH}_c} |g'(x)| > 0, \tag{B.2}$$

where $\bar{NH}_c$ is the closure of NH_c. Choosing $\epsilon = (1/2)\min(\epsilon_1, \epsilon_2)$ fulfils our purpose. To see this, choose $\tilde{g}$ such that $||g - \tilde{g}|| < \epsilon$. For $x \notin \bigcup_c NH_c$, $\tilde{g}$ is bounded away from zero by (B.1). In NH_c, there must be

exactly one x such that $\tilde{g}(x) = 0$: Because NH are mutually disjoint, $sgn(g(x_c - \delta)) \neq sgn(g(x_c + \delta))$, by (B.1) again $sgn(g(x_c - \delta)) = sgn(\tilde{g}(x_c - \delta))$ and $sgn(g(x_c + \delta)) = sgn(\tilde{g}(x_c + \delta))$, and finally the sign of $\tilde{g}'$ is constantly equal to $sgn(g'(x_c))$ in NH_c by (B.2).

The assertion for Z_ρ follows now from the first part of this proof, combined with Proposition 2.1, if we can choose ρ independent of $\tilde{g}$ such that Proposition 2.1 applies. Sufficient for this is ρ that separates roots. Choosing $\rho = \epsilon$ accomplishes this. By (B.1), L_ρ will separate the NH_c, and by the previous argument each of the NH_c will contain exactly one root of $\tilde{g}$. □

Proof of Theorem 2.2: We use Z^1, Z^2, Z^3 to denote a sequence of approximations to $\widehat{Z}$. Write $Z^1 =^A Z^2$, if $a_n Z^1 - b_n$ and $a_n Z^2 - b_n$ have the same non-degenerate distributional limit for some non-random sequences a_n and b_n. In particular, as long as such sequences exist that guarantee convergence to a non-degenerate limit, this is implied by equality up to a remainder, which is asymptotically negligible under the given sequence of experiments (i.e. $Z^1 =^A Z^2$ if $Z^1 - Z^2 = o_p(Z^2)$).

1. Approximation of $\widehat{g}$ with g

$$\widehat{Z} =^A Z_\rho(g, \widehat{g}').$$

The remainder of this approximation is given by

$$\int (L_\rho(g) - L_\rho(\widehat{g}))|\widehat{g}'|.$$

Negligibility of this remainder follows if we can show uniform convergence of $\widehat{g}$ at a rate faster than ρ under our sequence of experiments. Under the given sequence of experiments, the variance of $\widehat{g}(x)$ is of order $r_n^2/(n\tau)$ – this follows from the Bahadur expansion. Because we have assumed $r_n = (n\tau^3)^{1/2}$, we obtain $\mathrm{Var}(\widehat{g}(x)) = O(\tau^2) = o(1)$, which implies pointwise convergence. Pointwise convergence at rate τ implies uniform convergence at the slightly slower rate $\sqrt{\log(n)} \cdot \tau$, which is faster than ρ by our assumptions on τ and ρ; for background on uniform convergence of kernel estimators, see, e.g. Appendix A.1 of Armstrong (2014).

The fact that $\sup_x |\widehat{g}(x) - g(x)| = O_p(\sqrt{\log(n)} \cdot \tau)$ implies that the remainder $\int (L_\rho(g) - L_\rho(\widehat{g}))|\widehat{g}'|$ is of the same order. To see this, note that $\widehat{g}'$ is $O_p(1)$, so that the remainder is of the same order as $\int (L_\rho(g) - L_\rho(\widehat{g}))$. The integrand of this expression is non-zero only in a neighbourhood of size of order ρ of the roots of g; the difference $|L_\rho(g) - L_\rho(\widehat{g})|$ is of order $|\widehat{g} - g|/\rho$ because L_ρ is Lipschitz with constant C/ρ, so that the claim follows.

Thus, we have shown that the remainder $\int (L_\rho(g) - L_\rho(\widehat{g}))|\widehat{g}'|$ is of order $\sqrt{\log(n)} \cdot \tau$; this is smaller than the order of the leading term of $\widehat{Z}$, which we show to be ρ/τ. The remainder is thus asymptotically negligible.

From the approximation $\widehat{Z} =^A Z_\rho(g, \widehat{g}')$ we immediately obtain $\widehat{Z} =^A 0$ if $Z = 0$, because in that case $Z_\rho(g, \widehat{g}') = 0$ for ρ small enough. The claim of Theorem 2.2 is thus trivially satisfied for the case $Z = 0$, and we assume $Z > 0$ for the rest of this proof.

2. Approximation of $\widehat{g}'$ by the Bahadur expansion.

$$Z_\rho(g, \widehat{g}') = {}^A \int L_\rho(g(x))|g'(x) - f^{-1}(x)s_n^{-1}(x)I_n(x)$$
$$\times \frac{1}{n}\sum_i K_\tau(X_i - x)\phi(Y_i - g(x) - g'(x)(X_i - x))\left(\frac{X_i - x}{\nu_2 \tau^2}\right)|dx =: Z^1.$$

The absolute value of the remainder of this approximation is less than or equal to

$$\int L_\rho(g)|R|,$$

where R is the remainder of the Bahadur expansion. Negligibility of the remainder of the approximation is a consequence of the assumption that the remainder of the Bahadur expansion is negligible (i.e. $R = o_p((\widehat{g}, \widehat{g'}) - (g, g'))$ uniformly in x).

3. Restriction to one root at 0 and Taylor approximations

Assume that $g(0) = 0$ and $g(x) \neq 0$ for $x \neq 0$ (i.e. $Z = 1$). This is without loss of generality, because the integral for the general case is simply a sum of the independent integrals in a neighbourhood of each root.

Now define $c = g'(0)$, $w = -f^{-1}(0)s^{-1}(0)(1/\nu_2)$, $\phi_i = \phi(Y_i - g(x) - g'(x)(X_i - x))$ and $\tilde{K}_\tau(d) = K_\tau(d)(d/\tau)$.

By replacing g with $g'(0)x$ in $L_\rho(g(x))$ and replacing $-f^{-1}(x)s^{-1}(x)(1/\nu_2)$ with w, both justified by smoothness and $\rho \to 0$, as well as $I_n(x) \to 1$ uniformly, we obtain

$$\begin{aligned} Z^1 &= {}^A \int L_\rho(cx)|g'(0) - f^{-1}(0)s^{-1}(0)\frac{1}{\nu_2\tau^2}\frac{1}{n}\sum_i K_\tau(X_i - x)(X_i - x)\phi_i|dx \\ &= \int L_\rho(cx)|c + w\frac{r_n}{\tau}E_n[\tilde{K}_\tau(X_i - x)\phi_i]|dx = Z^2, \end{aligned}$$

where we use E_n to denote the sample average. The absolute value of the remainder of this approximation is less than or equal to

$$\int |L_\rho(g) - L_\rho(cx)||g' - \sum| + \int L_\rho(cx)|f^{-1}(x)s^{-1}(x)I_n(x)\frac{1}{\nu_2\tau^2} - w||E_n|.$$

Here, we use $\sum$ and E_n as shorthand for the sum and sample average of the previous display. Both terms in this expression go to 0 as $\rho \to 0$. We can assume furthermore that

$$X_i \sim^{\text{i.i.d.}} Uni([-\rho/c, \rho/c]),$$

conditional on falling in this interval, and that

$$\phi_i \sim^{\text{i.i.d.}} \phi(e)|X = 0.$$

These assumptions are justified by another Taylor approximation, this time of the distribution functions $F_X(x) = F_X(0) + f_X(0)X + o(X)$ and $F_{\phi|X}(\phi|X) = F_{\phi|X}(\phi|0) + O(X)$, assuming both distribution functions to be differentiable at 0. To see that this approximation is justified, note that distributional convergence to the same limit is equivalent to convergence of the expectations of any Lipschitz continuous bounded function of the statistics to the same limit. The difference in expectations between a function h of Z^2 and of its approximation using conditionally uniform X and i.i.d. ϕ is given by

$$\int h(Z^2)\prod_i (f_X(X^i)f_{\phi|X}(\phi^i|X^i) - f_X(0)f_{\phi|X}(\phi^i|0)).$$

This integral goes to 0 because the support of $h(Z^2)$ in X is a neighbourhood of 0 shrinking to 0.

4. Partitioning the range of integration

Partition $[-\rho/c, \rho/c]$ into subintervals $[t_j, t_{j+1}]$, $j = 1, \ldots, \lfloor \rho/\tau \rfloor$ with $t_{i+1} - t_i = 2\tau$. Then

$$Z^2 =^A \sum_{j=1}^{\lfloor \rho/c\tau \rfloor} L_\rho(ct_j)\xi_j = Z^3$$

with

$$\xi_j = \int_{t_j}^{t_{j+1}} |c + w\frac{r_n}{\tau^2} E_n[\tilde{K}_\tau(X_i - x)\phi_i]|dx.$$

The remainder of this approximation is given by

$$\int (L_\rho(cx) - L_\rho(c(\max_{t_j<x} t_j)))|c + w\frac{r_n}{h^2} E_n|.$$

This approximation is warranted by Lipschitz continuity of L_ρ with a Lipschitz constant of order $1/\rho^2$, and by $\tau/\rho^2 \to 0$.

5. Poisson approximation

The following argument essentially replaces the number of X falling into the interval $[-\rho/c, \rho/c]$, which is approximately distributed $Bin(n, 2f(0)\rho/c)$, with a Poisson random variable with parameter $2nf(0)\rho/c$; the distribution of everything else conditional on this number remains the same.

Let n_j be distributed i.i.d. $Poisson(2n\tau f(0))$ for $j = 1, \ldots, \lfloor \rho/\tau \rfloor$. This is an approximation to the number of X falling into the bin $[t_j, t_{j+1}]$. Draw $X_{jl} \sim^{\text{i.i.d.}} Uni([t_j, t_{j+1}])$ and $\phi_{jl} \sim^{\text{i.i.d.}} \phi(e)|X = 0$ for $j = 1, \ldots, \lfloor \rho/\tau \rfloor$ and $l = 1, \ldots, n_j$. Now define

$$\pi_j = \int_{t_j}^{t_{j+1}} |c + w\frac{r_n}{n\tau} \sum_{k=j-1}^{j+1} \sum_{l=1}^{n_k} [\tilde{K}_\tau(X_{jl} - x)\phi_{jl}]|dx.$$

Then

$$Z^3 =^A \sum_{j=1}^{\lfloor \rho/c\tau \rfloor} L_\rho(ct_j)\pi_j$$

where π_j are identically distributed and π_j is independent of π_k for $|j - k| \geq 2$.

Conditional on $\tilde{n} := \sum_j n_j$, the equality is exact. The exact distribution of the number of observations falling in the interval $[-\rho/c, \rho/c]$, corresponding to $\tilde{n}$, would be given by

$$\frac{(2n(\rho/c)f(0))^{\tilde{n}}}{\tilde{n}!} \frac{n!}{n^{\tilde{n}}(n - \tilde{n})!} (1 - 2(\rho/c)f(0))^{(n-\tilde{n})}.$$

The Poisson approximation sets the latter part of this expression to a constant in $\tilde{n}$. This is justified by the usual arguments deriving the Poisson distribution as a limit of Binomial distributions. The approximation of Z^3 follows by an argument similar to the second part of step 3 in this proof, once we note that the multinomial probability mass function converges uniformly.

6. Moments of the integrals over the subintervals

(a) $E[\pi_j] = \tau\mu_1 + o(\tau)$.
(b) $E[\pi_j^2] = \tau^2\mu_2 + o(\tau^2)$.
(c) $E[\pi_j\pi_{j+1}] = \tau^2\mu_{11} + o(\tau^2)$.
(d) $E[\pi_j^3] = \tau^3\mu_2 + o(\tau^3)$.

These equations follow from noting first pointwise convergence to normality of

$$\Gamma(x) = w\frac{r_n}{n\tau} \sum_{k=j-1}^{j+1} \sum_{l=1}^{n_k} (\tilde{K}_\tau(X_{jl} - x)\phi_{jl}) \xrightarrow{d} N(0, v)$$

under our sequence of experiments. This is the point where the rate r_n matters:

$$\begin{aligned}\Gamma(x) &= w\frac{\tau^{1/2}}{n^{1/2}}\sum_{k=j-1}^{j+1}\sum_{l=1}^{n_k}(\tilde{K}_\tau(X_{jl}-x)\phi_{jl})\\ &\sim w\frac{1}{(n\tau)^{1/2}}\sum_{l=1}^{(n_{j-1}+n_j+n_{j+1})}(K(\zeta_l)\zeta_l\phi_l)\\ &= w\Big(\frac{n_{j-1}+n_j+n_{j+1}}{n\tau}\Big)^{1/2}\Big(\frac{1}{n_{j-1}+n_j+n_{j+1}}\Big)^{1/2}\sum_{l=1}^{(n_{j-1}+n_j+n_{j+1})}(K(\zeta_l)\zeta_l\phi_l).\end{aligned}$$

Here, ζ_j are i.i.d. $Uni[-3,3]$. Now asymptotic normality follows by noting

$$\Big(\frac{n_{j-1}+n_j+n_{j+1}}{n\tau}\Big)\xrightarrow{p}6f(0),$$

$(n_{j-1}+n_j+n_{j+1})\xrightarrow{p}\infty$ and $E[\phi_l|X_l]=0$. Similarly

$$\begin{pmatrix}\Gamma(x_1)\\ \Gamma(x_1+\tau\delta)\end{pmatrix}\xrightarrow{d}N\Big(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}v & corr(|\delta|)\cdot v\\ corr(|\delta|)\cdot v & v\end{pmatrix}\Big).$$

Second, a change of the order of integration and the limit in n delivers the claims, where this change of order is justifiable by the dominated convergence theorem. For instance,

$$\begin{aligned}\lim(E[\pi_j^2]/\tau^2) &= 4\lim E[\int_{[0,1]^2}|(c+\Gamma(t_j+2\tau\delta_1))(c+\Gamma(t_j+2\tau\delta_2))|d\delta_1 d\delta_2]\\ &= 4\int_{[0,1]^2}\lim E[|(c+\Gamma(t_j+2\tau\delta_1))(c+\Gamma(t_j+2\tau\delta_2))|]d\delta_1 d\delta_2.\end{aligned}$$

7. Central limit theorem applied to the sum of integrals over the subintervals

Now apply a central limit theorem for m-dependent sequences to the sum of integrals. For a definition of m-dependence, see Hoeffding and Robbins (1994). Note that $L_\rho(ct_j)\pi_j$ is an m-dependent sequence with $m=1$. We have

$$\begin{aligned}&\mathrm{Var}(\sqrt{\frac{\rho}{\tau}}\sum_{j=1}^{\lfloor\rho/c\tau\rfloor}L_\rho(ct_j)\pi_j)\\ &= \frac{\rho}{\tau}\Big(\sum_j L_\rho^2(ct_j)\mathrm{Var}(\pi_j)+\sum_j L_\rho(ct_j)(L_\rho(ct_{j-1})+L_\rho(ct_{j+1}))\mathrm{Cov}(\pi_j,\pi_{j+1})\Big)\\ &\approx \Big(\frac{\rho}{\tau}\Big)\Big(\frac{c}{\tau}\Big)\int_{-\rho/c}^{\rho/c}L_\rho^2(cu)\tau^2(\mu_2+2\mu_{11}-3\mu_1^2)du\\ &= c(\mu_2+2\mu_{11}-3\mu_1^2)\int_{-1}^{1}L_1^2(cu)du.\end{aligned}$$

Asymptotic normality for $\sqrt{(\rho/\tau)}(Z^3-E[Z^3])$ follows, and by $\widehat{Z}\overset{A}{=}Z^3$, the same holds for $\sqrt{(\rho/\tau)}(\widehat{Z}-E[\widehat{Z}])$. Furthermore, $E[Z^3]=O(1)$, and hence so is $E[\widehat{Z}]$. □

Proof of Theorem 2.3: Fix one of the roots x_0 of g. By the arguments of the proof of Theorem 2.2, $\partial/\partial x\ \widehat{g}(x)$ (not to be confused with $\widehat{g'}(x)$) converges to a non-degenerate normal distribution for all x. In particular,

$$\liminf P(sgn(\partial/\partial x\ \widehat{g}(x_0)) \neq sgn(g'(x_0))) > 0.$$

By uniform convergence in levels of $\widehat{g}$ and the intermediate value theorem (compare also Figure 2),

$$P(Z(\hat{g}) > Z(g)) \geq P(sgn(\partial/\partial x\ \widehat{g}(x_0)) \neq sgn(g'(x_0))).$$

This proofs the first claim. The second claim now immediately follows from $\rho/\tau \to \infty$. □

Proof of Theorem 3.1 (SKETCH): We approximate $M(a, b, x, w_1)$ by a criterion function that has the form of (2.3) (i.e. a local weighted average over the empirical distribution of some objective function). Based on this approximation, we can then again apply the results of Kong et al. (2010). Newey (1994) provides a set of results that facilitate such approximations of partial means. In particular, Lemma 5.4 in Newey (1994) allows derivation of the required approximation by replacing the outer sum over j in (3.2) with an expectation, and by linearizing the fraction inside. The first replacement is asymptotically warranted because the variation created by averaging over the empirical distribution is of order $1/\sqrt{n}$ and is hence dominated by the variation in the non-parametric component. The second replacement follows from differentiability and requires, in particular, that the denominator of the fraction be asymptotically bounded away from zero. This is guaranteed by the requirement that W_2 has full conditional support given (X, W_1). Formally, Lemma 5.4 in Newey (1994) gives

$$M(a, b, x, w_1) - E_{W_2}[E_{m|X,W}[m|X = x, W_1 = w_1, W_2]] = \tilde{M}(a, b, x, w_1) + o_p(\tilde{M}(a, b, x, w_1)),$$

where

$$\tilde{M}(a, b, x, w_1) := \frac{1}{n} \sum_j \Bigg(K_\tau(X_j - x, W_{1j} - w_1) \times \frac{m(Y_j - a - b(X_j - x)) - E[m(Y_j - a - b(X_j - x))|X_j, W_j]}{f_{X,W_1|W_2}(X_j, W_{1j}|W_{2j})} \Bigg). \tag{B.3}$$

This approximation of the objective function has the general form assumed in Kong et al. (2010) if we set

$$\tilde{m}(Y, X, W, a, b, x) := \frac{m(Y - a - b(X - x)) - E[m(Y - a - b(X - x))|X, W]}{f_{X,W_1|W_2}(X, W_1|W_2)}, \tag{B.4}$$

providing us with the desired Bahadur expansion. Choosing the appropriate sequence of experiments, from here on the entire proof and result of Theorem 2.2 go through unchanged. If $W_1 \neq$ const, the rates have to be adapted as follows. The number of observations within each rectangle of size τ^d goes to ∞ if $n\tau^d \to \infty$. Finally, the variance of $\widehat{g'}$ converges iff $r_n = O(n\tau^{(2+d)})^{1/2}$. □

Proof of Theorem 3.2: The proof requires the following modifications relative to the one-dimensional case. Assumption 2.2 is still applicable, where the only difference in the d-dimensional case is that (2.6) has to be multiplied by $1/\tau^{d-1}$. For $\widehat{g'}$ to have a pointwise non-degenerate distributional limit, we have to choose the rate r_n to equal $(n\tau^{2+d})^{1/2}$, which is slower for higher d. To see this, note that $\text{Var}(\widehat{g'}) = O((r_n^2)/(n\tau^{2+d}))$. Here, L_ρ is Lipschitz continuous of order $\rho^{-(1+d)}$, so that we require $\tau/\rho^{d+1} \to 0$ for step 4 of the proof of Theorem 2.2. The range of integration has to be partitioned into rectangular subranges of area τ^d instead of intervals of length τ. There will be approximately const $\cdot(\rho/\tau)^d$ such subintegrals. The variance of the

integral of $|\widehat{g'}|$ over each of these subranges will be of order τ^{2d}, similarly for expectations and covariances. This yields a variance of $\widehat{Z}$ of $O((\tau/\rho)^d)$; see step 7 of the proof of Theorem 2.2. □

Proof of Theorem 3.3: By (3.14) and (3.16), it is sufficient to show that $r_n \cdot \widehat{g'^{-1}_1}(g_{2,n}(\bar{\sigma}_1, s_2), s_1)$ and $r_n \cdot \widehat{g'^{-1}_2}(\bar{\sigma}_1, s_1))$ converge jointly in distribution, while $r_n \cdot \widehat{g}(\bar{\sigma}_1, s)$, as well as $\widehat{\sigma}$, converge in probability. These claims follow as before if we combine the convergence of $r_n g_n$ from display (3.22) with Bahadur expansion (2.6) for $\widehat{g'^{-1}_2}$ and $\widehat{g'^{-1}_1}$, where the latter are evaluated at $\bar{\sigma}_{2,n}$, which is not constant but converges. □

Proof of Lemma 3.1: By definition of conditional quantiles, $F^{\Delta X|X}(Q^{\Delta X|X}(q|X)|X) = q$. Differentiating this with respect to X gives

$$\frac{\partial}{\partial X} Q^{\Delta X|X}(q|X) = -\frac{(\partial/\partial X) F^{\Delta X|X}(Q|X)}{f^{\Delta X|X}(Q|X)}. \tag{B.5}$$

The differential in the numerator has two components, one due to the structural relation between ΔX and X (i.e. the derivative with respect to the argument X of $d(X, \epsilon)$), and one due to the stochastic dependence of X and ϵ:

$$\begin{aligned}\frac{\partial}{\partial X} F^{\Delta X|X}(Q|X) = {} & E[g_X \cdot f^{\Delta X|g_X, X}(Q|g_X, X)|X] \\ & + \frac{\partial}{\partial X} \mathbb{P}(g(X', \epsilon) \leq Q|X)|_{X'=X}.\end{aligned}$$

This can be seen as follows: We can decompose the derivative according to

$$\frac{\partial}{\partial X} F^{\Delta X|X}(Q|X) = \Big[\frac{\partial}{\partial X'} + \frac{\partial}{\partial X}\Big] \mathbb{P}(g(X', \epsilon) \leq Q|X)|_{X'=X}.$$

To simplify the first derivative, note that by iterated expectations

$$\mathbb{P}(g(X', \epsilon) \leq Q|X) = E[F(g(X', \epsilon)|X, g_X)|X].$$

Differentiating this with respect to X' gives

$$E[g_X \cdot f^{\Delta X|g_X, X}(Q|g_X, X)|X].$$

The claim now is immediate. □

Proof of Proposition 3.1: Because X and $X + \Delta X$ have their support in the interval $[0, 1]$, $Q^{\Delta X|X}(q|0) \geq 0$ and $Q^{\Delta X|X}(q|1) \leq 0$. Therefore, the unique root X of $Q^{\Delta X|X}(q|X)$ must be stable, $(\partial/\partial X) Q^{\Delta X|X}(q|X) \leq 0$. By Lemma 3.1 and Assumption 3.3, this implies that $E[g_X|\Delta X = Q, X] \leq 0$.

Finally, note that for all X where $(0, X)$ is in the support of $(\Delta X, X)$, there exists a q such that $Q^{\Delta X|X}(q|X) = 0$. □

Proof of Proposition 3.2: The claims are immediate, noting that $N_c = \bigcap_s [x_c^s, x_{c+1}^s]$ and similarly for P_c. Furthermore, $x_c^s \in S_c$ for all s, $c = 1, 3, \ldots$ and $x_c^s \in U_c$ for all s, $c = 2, 4, \ldots$. Next, $g(\cdot, e_{i,s}) < 0$ on $[x_c^s, x_{c+1}^s]$, $c = 1, 3, \ldots$ from which negativity on N_c follows, similarly for P_c.

Finally, under monotonicity of potential outcomes, assuming for simplicity differentiability of g,

$$\frac{\partial}{\partial e} x_c = -\frac{(\partial/\partial e) g}{(\partial/\partial x) g}.$$

The numerator is always positive by assumption, the denominator is negative for $c = 1, 3, \ldots$ and positive for $c = 2, 4, \ldots$ because we had assumed g positive for sufficiently small x. Hence, $(\partial/\partial e)x_c$ is positive for $c = 1, 3, \ldots$ and negative for $c = 2, 4, \ldots$. □

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publishers web site.

Online Appendix A1: Further Monte Carlo Results
Online Appendix A2: Empirical Results for all the Metropolitan Statistical Areas in th Sample
Online Appendix A3: Global Dynamics of Economic Growth
Replication Files