ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

A Computational Implementation of GMM

Jiti Gao and Han Hong

November 2014

Working Paper 24/14

A Computational Implementation of GMM *

Jiti Gao[a] and Han Hong[b]

November 26, 2014

Abstract

In this paper we study a statistical method of implementing quasi-Bayes estimators for nonlinear and nonseparable GMM models, that is motivated by the ideas proposed in Chernozhukov and Hong (2003) and Creel and Kristensen (2011) and that combines simulation with nonparametric regression in the computation of GMM models. We provide formal conditions under which frequentist inference is asymptotically valid and demonstrate the validity of the use of posterior quantiles. We also show that in this setting, local linear kernel regression methods have theoretical advantages over local constant kernel methods that are also reflected in finite sample simulation results. Our results also apply to both exactly and over identified models. These estimators do not need to rely on numerical optimization or Markov Chain Monte Carlo simulations. They provide an effective complement to the classical M-estimators and to MCMC methods, and can be applied to both likelihood based models and method of moment based models.

Keywords: M-estimators, Monte Carlo Markov Chain methods, Nonparametric Regressions.

JEL Classification: C12, C15, C22, C52.

*We thank helpful comments by Victor Chernozhukov, Xiaohong Chen, Hide Ichimura, Michael Jansson, Sung Jae Jun, Dennis Kristensen, Joris Pinkse, Jim Powell and participants in various conferences and seminars, and Tingting Cheng for able research assistance. The first author was supported by an Australian Research Council Professorial Fellowship Award: DP1096374 and an Australian Research Council Discovery Projects Scheme under Grant number: DP130104229. The second author acknowledges financial support by the National Science Foundation (SES 1164589) and both the Department of Economics and SIEPR at Stanford.

[a] Monash University

[b] Stanford University

1 Introduction

Estimation of nonlinear models often involves difficult numerical optimization that might also be used in conjuction with simulation methods. Examples include maximum simulated likelihood, simulated method of moments, and efficient method of moments (Gallant and Tauchen (1996), Gourieroux, Monfort, and Renault (1993), Pakes and Pollard (1989)). Despite extensive efforts, see notably Andrews (1997), the problem of extremum computation remains a formidable impediment in these mentioned applications.

A computationally attractive alternative to M-estimators is the Markov Chain Monte Carlo methods (MCMC) which draws from the posterior distribution of the parameters given the data sample, and can be applied to both likelihood based and non-likelihood based models (Chernozhukov and Hong (2003)). However, both the maximum likelihood method and the MCMC method requires the evaluation of the likelihood function at each iteration of the Markov chain, which can be difficult to obtain even numerically.

A brilliant recent contribution by Creel and Kristensen (2011) demonstrates that it is possible to estimate highly complex and nonlinear parametric models by combining model simulation with nonparametric regressions. Their idea is based on simulating for models even when the likelihood function is difficult or infeasible to evaluate. This method only requires the ability to simulate from the model for each parameter value θ to compute a fixed dimensional summary statistics $\mathbf{T}_n$, and the ability of run a flexible (nonparametric) regression. The estimator is consistent, asymptotically normal, and asymptotically as efficient as a limited information maximum likelihood estimator. This is the first estimator that does not require either optimization, or MCMC, or the complex evaluation of the likelihood function.

In this paper we generalize the remarkable work by Creel and Kristensen (2011) in several important dimensions. First, we generalize these approximate Bayesian estimators to non-separable method of moment (GMM) models. Second, we formally demonstrate the validity of using posterior quantiles for frequentist inference under suitable regularity conditions. Third, by formalizing the interaction between simulation variance and sampling variance we discover importance differences between local linear (and generally local polynomial) regres-

sions and kernel regressions. Higher order local polynomial methods reduce both variance and bias, while higher order kernels only serve to reduce bias in these regressions. We also discuss issues related to different simulation sample sizes and misspecification. Although identifying θ using the zero root of the moment condition appears inherently local in nature. In subsequent work we plan to study the sieve based method (Chen (2007)) for estimating the posterior distribution, which has substantial computational advantage when used in combination with bootstrap inference for models without the information matrix equality.

Recently, Gentzkow and Shapiro (2013) suggest regressing the influence function of parameter estimates on moment conditions. Our goal differs in that we are providing a tool for parameter estimation and are not concerned about the identifying relation between moments and parameters. We also use nonparametric regressions instead of linear regressions. Furthermore, Jun, Pinkse, and Wan (2009) and Jun, Pinkse, and Wan (2011) develop generalized Laplace type estimators, and allow for nonstandard limiting distributions. Gallant and Hong (2007) used the Laplace transformation to recover the latent distribution of the pricing kernels.

In Section 2 below, we describe the estimator and inference method, starting with revisiting the Bayesian indirect inference estimator of Creel and Kristensen (2011), and proceeding to a generalization to the GMM model.

2 Motivation and the estimators

M-estimators, which are typically defined by $\tilde{\theta} = \arg\max_{\theta\in\Theta} \hat{Q}_n(\theta)$ for a sample objective function and parameter space $\Theta \in R^k$, can be difficult to compute. For m possibly different from the sample size n, a variant of the quasi-Bayesian estimator described in Chernozhukov and Hong (2003) replaces the M-estimator $\tilde{\theta}$ with a quasi-posterior mean estimator

$$\bar{\theta} = \frac{\int \theta\pi(\theta)\exp\left(m\hat{Q}_n(\theta)\right)d\theta}{\int \pi(\theta)\exp\left(m\hat{Q}_n(\theta)\right)d\theta}. \tag{1}$$

To implement $\bar{\theta}$, Chernozhukov and Hong (2003) suggest approximating $\bar{\theta}$ using Monte Carlo Markov Chain (MCMC), which itself can be computationally intensive and also requires the ability to evaluate $\hat{Q}_n(\theta)$. Formalizing the convergence of the MCMC chain to the

posterior distribution is also difficult. See for example Belloni and Chernozhukov (2009). Alternative estimators, which require neither optimization nor MCMC, can be desirable for both likelihood based models and memthod of moment models.

In parametric models, an insightful result in Creel and Kristensen (2011) shows that when $m = n$ and when $\hat{Q}(\theta)$ equals the log likelihood of a set of summary statistics T_n of dimension d, the Bayes estimator $\bar{\theta} = E(\theta|T_n; \pi(\theta), f(T_n|\theta))$ in (1) can be approximated through a combination of simulation and nonparametric regressions. For $s = 1, \ldots, S$, draw θ_s from $\pi(\theta^s)$, and then draw T_n^s from the parametric model $f(T_n^s|\theta^s)$. Given $\theta^s, T_n^s, s = 1, \ldots, S$, Creel and Kristensen (2011) propose to approximate $\bar{\theta}$ using a kernel estimator

$$\hat{\theta} = \sum_{s=1}^{S} \theta^s \kappa\left(\frac{T_n^s - T_n}{h}\right) / \sum_{s=1}^{S} \kappa\left(\frac{T_n^s - T_n}{h}\right).$$

In principle, the simulation sample size, denoted m, can differ from the sample size n.

This insightful idea will be generalized in several substantial directions. Most importantly, we show that this sampling and regression framework also applies to general nonseparable method of moment models. Consider a set of d dimensional sample moments of the form of $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i, \theta)$. For these models, Chernozhukov and Hong (2003) suggest implementing the Laplace transformation using $\hat{Q}_n(\theta) = -\frac{1}{2}\hat{g}(\theta)' \hat{W}(\theta) \hat{g}(\theta)$, where $\hat{W}(\theta)$ is a possibly data and parameter dependent weighting matrix. In the following we will focus on the case when $\hat{W}(\theta)$ is the inverse of a consistent estimate of the asymptotic variance of $\hat{g}(\theta)$. For example, with i.i.d. data, $\hat{W}(\theta) = \hat{\Sigma}(\theta)^{-1}$, where $\hat{\Sigma}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i; \theta) g(Z_i; \theta)'$. In this case, equation (1) takes the form of

$$\bar{\theta} = \frac{\int \theta \pi(\theta) \exp\left(-\frac{1}{2} m \hat{g}(\theta)' \hat{W}(\theta) \hat{g}(\theta)\right) d\theta}{\int \pi(\theta) \exp\left(-\frac{1}{2} m \hat{g}(\theta)' \hat{W}(\theta) \hat{g}(\theta)\right) d\theta}. \tag{2}$$

Consider now a random variable Y_m, which, given θ and the data, follows a d dimensional multivariate normal distribution with mean vector $\hat{g}(\theta)$ and variance-covariance matrix $\frac{1}{m}\hat{\Sigma}(\theta)$: $N\left(\hat{g}(\theta), \frac{1}{m}\hat{\Sigma}(\theta)\right)$. When θ is drawn from the prior density $\pi(\theta)$, the posterior density of θ given $Y_m = y$ is

$$f(\theta|Y_m = y) \propto \pi(\theta) \det\left(\hat{\Sigma}(\theta)\right)^{-\frac{1}{2}} \exp\left(-\frac{m}{2}(\hat{g}(\theta) - y)' \hat{W}(\theta)(\hat{g}(\theta) - y)\right).$$

Then we note that a slight variation of(2) can be written as $\bar{\theta} = \bar{\theta}(y)\Big|_{y=0} \equiv \bar{\theta}(0)$, where

$$\begin{aligned}\bar{\theta}(y) =& E(\theta|Y_m = y) \\ =& \frac{\int \theta\pi(\theta) \det\left(\hat{\Sigma}(\theta)\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}m(\hat{g}(\theta) - y)'\hat{\Sigma}(\theta)^{-1}(\hat{g}(\theta) - y)\right) d\theta}{\int \pi(\theta) \det\left(\hat{\Sigma}(\theta)\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}m(\hat{g}(\theta) - y)'\hat{\Sigma}(\theta)^{-1}(\hat{g}(\theta) - y)\right) d\theta.}\end{aligned}$$

This interpretation suggests the following simulation based estimator:

- Draw $\theta^s, s = 1, \ldots, S$ from $\pi(\theta)$. For each θ^s, compute $\hat{g}(\theta^s)$.
- Next draw Y_m^s from $N\left(\hat{g}(\theta^s), \frac{1}{m}W(\theta^s)^{-1}\right)$: $Y_m^s = \hat{g}(\theta^s) + \frac{1}{\sqrt{m}}W(\theta^s)^{-1/2}\xi$, for $\xi \sim N(0, I_d)$.
- Using the simulated sample $\theta^s, Y_m^s, s = 1, \ldots, S$, run a flexible nonparametric regression of θ^s on Y_m^s, and impute the fitted value at $Y_m = 0$ as $\hat{\theta}$.

A variety of kernel or sieve based nonparametric regression methods can be used for this purpose. For example, the kernel method is used in Creel and Kristensen (2011), and makes the best use of a kernel function $\kappa(\cdot)$ and a bandwidth h to define

$$\hat{\theta} = \sum_{s=1}^{S} \theta^s \kappa\left(\frac{Y_m^s}{h}\right) / \sum_{s=1}^{S} \kappa\left(\frac{Y_m^s}{h}\right).$$

In this paper we focus instead on local linear kernel regressions, which is defined as $\hat{\theta} = \hat{a}$, where

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \left(\sum_{s=1}^{S} \begin{pmatrix} 1 \\ Y_m^s \end{pmatrix} \begin{pmatrix} 1 & Y_m^s \end{pmatrix} \kappa\left(\frac{Y_m^s}{h}\right)\right)^{-1} \left(\sum_{s=1}^{S} \begin{pmatrix} 1 \\ Y_m^s \end{pmatrix} \theta^s \kappa\left(\frac{Y_m^s}{h}\right)\right). \tag{3}$$

We find that it has theoretical advantage over local constant kernel methods. The results generalize readily to higher order polynomials and higher order kernels. In particular, while higher order kernels reduce bias, higher order polynomials reduce both bias and variance.

The previous setup mimics continuous updating GMM which chooses $\hat{W}(\theta)$ using the estimated variance of $\hat{g}(\theta)$ for each θ. The usual optimal two step GMM choice of the weighting matrix also applies. In step one, an ad hoc choice of $W(\theta)$ can be used, for example,

$W(\theta) = I$, to produce an initial $\sqrt{n}$ consistent but not necessarily efficient estimator, $\hat{\theta}_1$. Posterior quantile inference is invalid at this step. In the second step, choose $\hat{W}$ according to the $\hat{\Sigma}\left(\hat{\theta}_1\right)$, the estimated variance of $\hat{g}(\theta)$ evaluated at $\theta = \hat{\theta}_1$. The second step is efficient and provides valid posterior quantile inference.

The generalized nonseparable moment model setup is closely related to the original indirect likelihood method developed by Creel and Kristensen (2011), who analyzed the separable moment condition $\hat{g}(\theta) = T_n - t(\theta)$, where $t(\theta)$ is the probability limit of T_n under θ. When $t(\theta)$ does not admit an analytic expression, Creel and Kristensen (2011) replace it with a simulated version T_n^s from θ^s and use $Y_n^s = T_n - T_n^s$. This is tantamount to drawing Y_n^s from

$$\hat{g}(\theta^s) + \frac{1}{\sqrt{n}} W\left(\theta^2\right)^{-1/2} \xi_n^s = T_n - t(\theta^s) - (T_n^s - t(\theta^s)),$$

where ξ_n^s is approximately a standard normal random vector:

$$\xi_n^s = \Sigma(\theta^s)^{-\frac{1}{2}} \sqrt{n}\,(T_n^s - t(\theta^s)) \xrightarrow{d} N(0, I).$$

The unknown $t(\theta^s)$ cancels from the feasible moment condition, which is particularly appealing in parametric models with complex likelihood but that are feasible to simulate. By simulating from the likelihood, the method proposed by Creel and Kristensen (2011) can be computationally more efficient than directly using the GMM criterion function:

$$-\frac{n}{2}(T_n - t(\theta))' W(\theta)(T_n - t(\theta)). \tag{4}$$

Using (4) may involve more computation when (1) for each θ, $t(\theta)$ needs to be simulated with a simulation sample size larger than the observed sample size, for example $T_m(\theta)$, to avoid loss of efficiency; and (2) in an overidentified model where $\dim(T_n) > \dim(\theta)$, the optimal $W(\theta) = \Sigma(\theta)^{-1}$ also needs to be computed with a large number of simulations. Alternatively, one may proceed in two steps as described above. In the first step, an ad hoc choice of $W(\theta)$, e.g. $W(\theta) = I$, is used to produce an initial estimate $\hat{\theta}_1$. In step two, evaluate an estimate of the optimal weighting matrix with $W\left(\hat{\theta}_1\right) = \hat{\Sigma}\left(\hat{\theta}_1\right)^{-1}$. The method of combining simulation and nonparametric regression to approximate the *infeasible* estimators in (1) and (2) also falls into the high level framework of approximate estimators developed in Kristensen and Salanié (2010).

We also analyze the use of the entire posterior distribution of θ^s given Y_m^s for inference. We focus on a smooth scalar function of the parameters $\eta = \eta(\theta)$. Define the τ percentile of η given $Y_m = 0$, denoted $\bar{\eta}_\tau$, through the estimating equation of

$$\int_{-\infty}^{\bar{\eta}_\tau} \pi(\theta) f(Y_m^s = 0|\theta) d\theta / \int \pi(\theta) f(Y_m^s = 0|\theta) d\theta = \tau.$$

It will be shown that an asymptotically valid $1-\tau$th confidence interval for τ is given by

$$P\left(\eta_0 \in \left(\bar{\eta}_{1/2} + \sqrt{\frac{m}{n \wedge m}}\left(\bar{\eta}_{\tau/2} - \bar{\eta}_{1/2}\right), \bar{\eta}_{1/2} + \sqrt{\frac{m}{n \wedge m}}\left(\bar{\eta}_{1-\tau/2} - \bar{\eta}_{1/2}\right)\right)\right).$$

In the case when $n = m$, the confidence interval specializes to $\left(\hat{\eta}_{\tau/2}, \hat{\eta}_{1-\tau/2}\right)$. Other forms of intervals, such as symmetric intervals, can also be used.

Using the simulated sample $\theta^s, Y_m^s, s = 1, \ldots, S$, the posterior quantile $\bar{\theta}_\tau$ can be approximated by a nonparametric estimate of the conditional τth percentile of $\eta^s = \eta(\theta^s)$ given Y_m^s. For example, the kernel method solves for $\hat{\eta}_\tau$ from the estimating equation

$$\sum_{s=1}^{S} \left(1\left(\eta^s \leq \hat{\eta}_\tau\right) - \tau\right) \kappa\left(\frac{Y_m^s}{h}\right) = 0.$$

We also use instead local linear quantile regression to estimate $\bar{\eta}_\tau$, and define $\hat{\eta}_\tau = \hat{a}$ through

$$\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \rho_\tau\left(\eta^s - a - by^s\right) \kappa\left(\frac{Y_m^s}{h}\right).$$

In the above, $\rho_\tau(u) = (\tau - 1(u \leq 0))u$ is the Koenker and Bassett (1978) check function used in quantile regression. We provide conditions under which $\hat{\eta}_\tau$ is sufficiently close to $\bar{\eta}_\tau$ to provide asymptotically valid confidence intervals.

Heuristically, m may be taken to be ∞. In this case $Y^s = \hat{g}(\theta^s)$, where we attempt to estimate θ using the θ^s for which $Y^s = \hat{g}(\theta^s)$ is as close to zero as possible. This is implemented through a nonparametric local to zero regression of θ^s on Y^s. For example, using the kernel method,

$$\hat{\theta} = \sum_{s=1}^{S} \theta^s \kappa\left(\frac{\hat{g}(\theta^s)}{h}\right) / \sum_{s=1}^{S} \kappa\left(\frac{\hat{g}(\theta^s)}{h}\right).$$

This is analogous to a nonparametric regression in which there is no error term: $\epsilon \equiv 0$, $y = g(x)$. In this case, the variance of the estimator $\hat{g}(x)$ is solely due to the variation of

the conditinal expectation $g(x')$ for x' within the window centered at x controlled by kernel function and the bandwidth h. The conditional variance of y given x is not included.

This method can be shown to work fine for exactly identified models. However, for overidentified models, while local constant kernel methods can still be implemented, local linear or polynomial kernel methods involve possible multicolinearity among regressors $\hat{g}(\theta^s)$. For example, with $m = \infty$, local linear methods rely on (quadratic) nonlinearity of moment conditions to generate variations in the regressors to avoid collinearity. In the absense of this variation, the resulting colinearity creates indeterminancy of the predicted value at zero within a $1/\sqrt{n}$ neighborhood. Therefore, $\hat{\theta}$ may not converge to a well defined limiting distribution. There are two additional reasons to focus on $m = n$ even if it is possible to take $m = \infty$. The first is to be consistent with the MCMC setup in Chernozhukov and Hong (2003). The second is that in implementing the indirect likelihood inference in Kristensen and Shin (2012), computation cost is increasing in m. It is natural to take $n \leq m$. Consequently, we focus on developing results for $m = n$, and only discuss the extension to $m > n$ after presenting the main results.

Furthermore, when the model is overidentified with $d > t$, conditional on a realization of θ^s, the event that $\hat{g}(\theta^s) = t$ is not possible for almost all values of t (Lebesgue measure 1). In this case, the conditional distribution of $\theta|\hat{g}(\theta) = t$ is not defined for almost all t, including $t = 0$ for almost all realizations of $\hat{g}(\theta)$. On the other hand, for $m < \infty$, regardless of how large, the conditional distribution

$$\theta|Y \equiv \hat{g}(\theta) + \frac{\xi}{\sqrt{m}} = t$$

is always well defined for all t, as long as ξ has full support.

In the rest of the paper, Section 3 formally establishes an asymptotic theory. Section 4 discusses issues related to overidentification and misspecification, and alternative implementations. Section 5 reports the finite sample performance of the proposed methods based on a Monte Carlo simulation exercise. Section 6 contains an empirical application, and Section 7 concludes. Lemmas and proofs that are peripheral to the main results are collected in a technical addendum.

3 Asymptotic Distribution Theory

Often times the estimators $\bar{\theta}$ and $\bar{\eta}_\tau$ do not admit analytic expressions, and require the feasible implementation through the simulated versions of $\hat{\theta}$ and $\hat{\eta}_\tau$. We first state regularity conditions required for the *infeasible* estimators $\bar{\theta}$ and $\bar{\eta}_\tau$ to be consistent and asymptotically normal, which mirror the general results in Chernozhukov and Hong (2003). Subsequently, we will develop conditions on the number of simulations and the size of the simulated sample that are necessary for $\hat{\theta}$ and $\hat{\eta}_\tau$ to be asymptotically valid.

3.1 Asymptotic theory for the infeasible estimators

The first assumption is standard regarding the parameter space.

ASSUMPTION 1 The true parameter θ_0 belongs to the interior of a compact convex subset Θ of Euclidean space $\mathbb{R}^k$. The weighting function $\pi : \Theta \to \mathbb{R}_+$ is a continuous, uniformly positive density function.

Next, there are two sets of regularity assumptions. The first set of assumptions applies to a general objective function $\hat{Q}_n(\theta)$, and can be used for either the indirect inference model with $\hat{Q}_n(\theta) = \frac{1}{n}\log f(T_n|\theta)$, or for the GMM model with $\hat{Q}_n(\theta) = -\frac{1}{2}\hat{g}(\theta)' \hat{W} \hat{g}(\theta)$.

To state the asymptotic results we relate $\bar{\theta}(y)$ and $\bar{\eta}_\tau$ to the M estimator $\tilde{\theta}$. For this purpose, consider an indexed family of M estimators, defined by

$$\tilde{\theta}(y) = \arg\min_{\theta\in\Theta} \hat{Q}_n(y|\theta) = \frac{1}{n} log f(T_n + y|\theta) \quad \text{or} \quad \tilde{\theta}(y) = \arg\min_{\theta\in\Theta} (\hat{g}(\theta) - y)' \hat{W}(\theta) (\hat{g}(\theta) - y).$$

Also define a population analog of an indexed family of parameters:

$$\theta(y) = \arg\min_{\theta\in\Theta} Q(y|\theta) \equiv (t(\theta_0) - t(\theta) + y)' \Sigma(\theta) (t(\theta_0) - t(\theta) + y)$$

or

$$\theta(y) = \arg\min_{\theta\in\Theta} Q(y|\theta) \equiv (g(\theta) - y)' \Sigma(\theta)^{-1} (g(\theta) - y).$$

Then $\theta_0 = \theta(0)$ and $\tilde{\theta} = \tilde{\theta}(0)$.

ASSUMPTION 2 For any $\delta > 0$, there exists $\epsilon > 0$, such that

$$\liminf_{n\to\infty} P\left\{\sup_{y\in\mathcal{Y}} \sup_{|\theta-\theta_y|\geq\delta} \left(\hat{Q}(\theta, y) - \hat{Q}(\theta_y, y)\right) \leq -\epsilon\right\} = 1.$$

ASSUMPTION 3 There exist a family of quantities Δ_n^y, J_y, Ω, where $\sup_{y\in\mathcal{Y}} |\sqrt{n}\Delta_n^y| = O_P(1)$, and $\sqrt{n}\Omega^{-1/2}\Delta_n^0 \xrightarrow{d} N(0, I)$, such that if we write

$$R_n^y(\theta, \theta^*) = \hat{Q}_y(\theta) - \hat{Q}_y(\theta^*) - (\theta - \theta^*)' \Delta_n^y + \frac{1}{2}(\theta - \theta^*)'(J_y)(\theta - \theta^*),$$

then it holds that for any sequence of $\delta \to 0$

$$\sup_{y\in\mathcal{Y}} \sup_{|\theta-\theta^*|\leq\delta, \theta\in\mathcal{N}(\theta_0), \theta^*\in\mathcal{N}(\theta_0)} \frac{R_n^y(\theta, \theta^*)}{|\theta - \theta^*|^2 + |\theta - \theta^*|/\sqrt{n}} = o_P(1).$$

In addition for each $y \in \mathcal{Y}$, $Q_y(\theta)$ is twice differentiable at θ_y with uniformly nonsingular second derivative $H_y = -J_y$, so that $\inf_{y\in\mathcal{Y}} \inf_{|x|\neq 0} \frac{x'J_y x}{x'x} > 0$, and for any $\delta_n \to 0$,

$$\sup_{y\in\mathcal{Y}} \sup_{|\theta-\theta_y|\leq\delta_n} \frac{|Q_y(\theta) - Q_y(\theta_y) - \frac{1}{2}(\theta - \theta_y)' H_y (\theta - \theta_y)|}{|\theta - \theta_y|^2} = o(1).$$

For $\tilde{\theta} = \arg\max_\theta \hat{Q}_n(\theta)$. Under these assumptions, for large m, the normalized and recentered posterior distribution of the parameters $u = \sqrt{m}\left(\theta - \tilde{\theta}\right)$ is well approximated by a normal density, which takes the form of $p_\infty^*(u) = \sqrt{\frac{|J|}{(2\pi)^d}} \cdot \exp\left(-\frac{1}{2}u'Ju\right)$.

The second set of assumptions specializes to the GMM model.

ASSUMPTION 4 For any $\epsilon > 0$, there is $\delta > 0$ such that

$$\limsup_{n\to\infty} P\left\{\sup_{|\theta-\theta'|\leq\delta} \frac{\sqrt{n}|(\hat{g}(\theta) - \hat{g}(\theta')) - (g(\theta) - g(\theta'))|}{1 + \sqrt{n}|\theta - \theta'|} > \epsilon\right\} < \epsilon. \tag{5}$$

ASSUMPTION 5 For all $\theta \in \Theta$, $\sqrt{n}(\hat{g}(\theta) - g(\theta)) \xrightarrow{d} N(0, \Sigma(\theta))$. $\Sigma(\theta)$ is uniformly positive definite and finite on $\theta \in \Theta$. Furthermore, $\sup_{\theta\in\Theta} |\sqrt{n}(\hat{g}(\theta) - g(\theta))| = O_P(1)$.

ASSUMPTION 6 (1) $\hat{W}(\theta) = \hat{\Sigma}(\theta)^{-1}$ and $\sup_{\theta\in\Theta} |\hat{\Sigma}(\theta) - \Sigma(\theta)| = o_P(1)$;
(2) $\sup_{\theta\in\mathcal{N}(\theta_0)} \sqrt{n}|\hat{\Sigma}(\theta) - \Sigma(\theta)| = O_P(1)$.

Remark: Assumptions 3 and 4 are standard stochastic equicontinuity conditions. See for example, section 7.1 (pp 2185), Theorems 7.1 and 7.2 in Newey and McFadden (1994). Section 3.2 (pp 2255) in Andrews (1994) provides examples for which these conditions hold.

ASSUMPTION 7 Either one of the following conditions hold.

(a) The model is exactly identified $d = k$.

(b) $\sup_{y\in\mathcal{Y}} \frac{\sqrt{n}|(\hat{g}(\theta)-\hat{g}(\theta_y))-(g(\theta)-g(\theta_y))|}{|\theta-\theta_y|} = O_P(1)$.

(c) $\sup_{y\in\mathcal{Y}} \frac{\sqrt{n}|(\hat{g}(\theta)-\hat{g}(\theta_y))-(g(\theta)-g(\theta_y))|}{\sqrt{|\theta-\theta_y|}} = O_P(1)$, $nh^4 \to 0$.

(d) $nh^2 = O(1)$.

Remark: For exactly identified models we do not require assumption 7, which is only relevant for overidentified models. Assumption 7(b) applies to smooth models. For example, when $\hat{g}(\theta)$ is multiple times differentiable with bounded continuous derivatives, we expect

$$\hat{g}(\theta) - \hat{g}(\theta_y) - (g(\theta) - g(\theta_y)) = \left(\hat{G}(\theta_y) - G(\theta_y)\right)(\theta - \theta_y) + o_p\left(|\theta - \theta_y|^2\right).$$

On the other hand, Assumption 7(c) is used to handle nonsmooth models that involved indicator functions, such as overidentified quantile instrumental variables. See for example Kim and Pollard (1990), and Theorem 3.2.5 and pp 291 of Van der Vaart and Wellner (1996). In this case the proof is more involved, and we require $nh^{-1/4}$. Hence in this case we are not able yet to use higher order local polynomials to relax the condition on the bandwidth h and on the number of simulations S, although we allow for local linear regression.

The following theorem mirrors Chernozhukov and Hong (2003), and shows that uniformly in y, the posterior distribution of $\sqrt{m}\left(\theta - \bar{\theta}\right)$ given y accurately approximates the asymptotic distribution of $\sqrt{n}\left(\tilde{\theta}(y) - \theta(y)\right)$. This allows for valid asymptotic inference based on the quasi-posterior quantiles of $\bar{\eta}_\tau$.

THEOREM 1 Under Assumptions 1-2, or Assumptions 1 and 3-6, we have

1. $\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{J}^{-1}\Omega\mathcal{J}^{-1}\right)$ when Assumption 2 holds. The asymptotic variance is replaced by $J = G(\theta_0) W G(\theta_0)$ and $\Omega = G(\theta_0)' \Sigma(\theta_0)^{-1} G(\theta_0)$ for $G(\theta) = \frac{\partial}{\partial\theta} g(\theta)$ when Assumptions 3-6 hold.

2. Define the localized parameter $u = \sqrt{m}\left(\theta - \tilde{\theta}\right)$ with a posterior density:

$$p_n(u) = \frac{1}{\sqrt{m}} f_\theta\left(\tilde{\theta} + \frac{u}{\sqrt{m}} | Y_m = 0\right).$$

Then $p_n(u)$ converges to a multivariate normal density of the form

$$p_\infty(u) = \sqrt{\frac{|W(\theta_0)|}{(2\pi)^d}} \exp\left(-\frac{1}{2}u'W(\theta_0)u\right),$$

in a strong total variation norm:

$$\|p_n(h) - p_\infty(h)\|_{TMV(\alpha)} \equiv \int (1+|h|^\alpha)|p_n(h) - p_\infty(h)|dh = o_P(1).$$

3. Consequently, for any $\tau \in (0,1)$,

$$P\left(\eta_0 \in \left(\bar{\eta}_{1/2} + \sqrt{\frac{m}{n \wedge m}}\left(\bar{\eta}_{\tau/2} - \bar{\eta}_{1/2}\right), \bar{\eta}_{1/2} + \sqrt{\frac{m}{n \wedge m}}\left(\bar{\eta}_{1-\tau/2} - \bar{\eta}_{1/2}\right)\right)\right) \geq 1 - \tau + o(1).$$

The uniform convergence rate in part 1 of Theorem 1 is not needed for parts 2 and 3, but is needed in proving the feasible simulation estimators $\hat{\theta}$ and $\hat{\eta}_\tau$. In part 3 of Theorem 1, when $m \geq n$, converage is asymptotically exact. However, when $m << n$, the interval is only conservatively valid and the coverage rate converges to 1.

A local uniform version of the above theorem also holds as in the following lemma, which will be needed for the implementation of the nonparametric regressions.

LEMMA 1 Under the same conditions as in Theorem 1, uniformly in t in a shrinking neighbourhood of 0,

$$\sqrt{n}\left(\tilde{\theta}(t) - \theta(t)\right) = O_P(1) \quad \text{and} \quad \sqrt{m}\left(\bar{\theta}(t) - \tilde{\theta}(t)\right) = o_P(1).$$

3.2 Behavior of the marginal density

The large sample (in the number of simulations S) properties of nonparametric regressions of θ^s on $y_s = Y_m^s$ local to 0 depends on the properties of the conditional distribution of $f(\theta|Y=y)$, and on the marginal density of $f(Y=y)$ for y close to zero. Unlike conventional nonparametric regressions, both the marginal density of Y and the conditional variance (or conditional density in the quantile regression case) of θ given Y are sample–size dependent. We analyze each in turn.

Section 3.1 shows that $f(\theta|Y=y)$ concentrates on a $O\left(\frac{1}{\sqrt{n}}\right)$ neighbourhood of $\tilde{\theta}(y)$. Therefore we expect that $Var(\theta|y) = O(1/m)$ and that $f\left(\tilde{\theta}(y)|y\right) = O\left(m^{k/2}\right)$. The

convergence rates of the local linear and quantile estimators also depend on the behaviour of the marginal density $f_Y(y)$ when y is close to 0. It turns out that in an exactly identified model $f_Y(0) = O_p(1)$, while in an overidentified model where $d > k$, $f_Y(0) = O_P\left(m^{\frac{d-k}{2}}\right)$.

We first use a simple example to illustrate this.

EXAMPLE 1 To illustrate, consider the simple case of a vector of sample means. Let $\theta = \mu$, and $\hat{g}(\mu) = \mu - \bar{X}$. Let $\pi(\mu) = N(\mu_0, \Sigma_0)$. For $\xi \sim N(0,1)$, let

$$Y_m = \mu - \bar{X} + \frac{1}{\sqrt{m}}\Sigma^{1/2}\xi.$$

So that given μ, $Y_m \sim N\left(\mu - \bar{X}, \frac{1}{m}\Sigma\right)$. Then the posterior mean and variance are given by

$$E(\mu|Y_m = t) = \frac{\Sigma}{m}\left(\Sigma_0 + \frac{\Sigma}{m}\right)^{-1}\mu_0 + \Sigma_0\left(\Sigma_0 + \frac{\Sigma}{m}\right)^{-1}\left(\bar{X} + t\right) \overset{m\to\infty}{\longrightarrow} \bar{X} + t$$

and

$$Var(\mu|Y_m = t) = \Sigma_0\left(\Sigma_0 + \frac{1}{m}\Sigma\right)^{-1}\frac{\Sigma}{m} = O\left(\frac{1}{m}\right).$$

Given $\bar{X}$, the marginal density of Y_m is

$$N\left(\mu_0 - \bar{X}, \Sigma_0 + \frac{1}{m}\Sigma\right) = O_P(1),$$

whenever Σ_0 is nonsingular, as in the exact identification case of $d \equiv \dim(Y_m) = k \equiv \dim(\mu)$.

Suppose now $d > k = 1$, then for a scalar u_0 and σ_0^2, and for l being a $d \times 1$ vector of 1's, we can write $\mu_0 = u_0 l$ and $\Sigma_0 = \sigma_0^2 ll'$. The previous calculation can not be used when $m \to \infty$. Instead, note that

$$\left(\frac{\Sigma}{m} + \sigma_0^2 ll'\right)^{-1} = m\Sigma^{-1} - \frac{\sigma_0^2 m^2 \Sigma^{-1} ll' \Sigma^{-1}}{1 + \sigma_0^2 m l'\Sigma^{-1} l}.$$

In this case,

$$\begin{aligned} E(\mu|Y_m = t) &= \left(I - \frac{\sigma_0^2 m ll'\Sigma^{-1}}{1 + \sigma_0^2 m l'\Sigma^{-1} l}\right)u_0 l + \sigma_0^2 ll'\left(m\Sigma^{-1} - \frac{\sigma_0^2 m^2 \Sigma^{-1} ll'\Sigma^{-1}}{1 + \sigma_0^2 m l'\Sigma^{-1} l}\right)\left(\bar{X} + t\right) \\ &= \left(I - \frac{\sigma_0^2 ll'\Sigma^{-1}}{1/m + \sigma_0^2 l'\Sigma^{-1} l}\right)u_0 l + \frac{m\sigma_0^2 ll'\Sigma^{-1}}{1 + \sigma_0^2 m l'\Sigma^{-1} l}\left(\bar{X} + t\right). \end{aligned}$$

As $m \to \infty$, $E(\mu|Y_m = t) \to \frac{ll'\Sigma^{-1}}{l'\Sigma^{-1}l}\left(\bar{X} + t\right)$, which is the GLS estimator. Furthermore, (now interpret μ as a scalar):

$$Var(\mu|Y_m = t) = \sigma_0^2 - \sigma_0^4 l'\left(\Sigma_0 + \Sigma/m\right)^{-1} l = \sigma_0^2 \frac{1}{1 + \sigma^2 m l'\Sigma^{-1} l}.$$

The interesting part is the marginal density of Y_m at $t = 0$:

$$N\left(\bar{X} - u_0 l, \left(\frac{\Sigma}{m} + \sigma_0^2 ll'\right)\right)$$

as it becomes singular when $m \to \infty$. Let

$$\bar{X} - \mu_0 = \left(\bar{X}_1 - u_0\right) l + \left(0, \Delta/\sqrt{n}\right)' \quad \text{for} \quad \Delta = \sqrt{n}\left(\bar{X}_{-1} - \bar{X}_1\right)$$

so that $\Delta \sim N(0, \Omega)$ for some Ω when the model is correctly specified. Then the exponent of this density under correct specification becomes

$$\begin{aligned}
&-\frac{1}{2}\left(\bar{X} - \mu_0\right)' \left(\frac{\Sigma}{m} + \sigma_0^2 ll'\right)^{-1} \left(\bar{X} - \mu_0\right) \\
&= -\left(\bar{X}_1 - u_0\right)^2 l' \left(\frac{\Sigma}{m} + \sigma_0^2 ll'\right)^{-1} l - \left(0, \Delta/\sqrt{n}\right) \left(\frac{\Sigma}{m} + \sigma_0^2 ll'\right)^{-1} \left(0, \Delta/\sqrt{n}\right)' \\
&= -\left(\bar{X}_1 - u_0\right)^2 \frac{ml'\Sigma^{-1}l}{1 + \sigma_0^2 ml'\Sigma^{-1}l} - \left(0, \Delta/\sqrt{n}\right) \left(m\Sigma^{-1} - \frac{\sigma_0^2 m^2 \Sigma^{-1} ll' \Sigma^{-1}}{1 + \sigma_0^2 ml'\Sigma^{-1}l}\right) \left(0, \Delta/\sqrt{n}\right)'.
\end{aligned}$$

When $m = O(n)$, this is $O_P(1)$. However, when $m >> n$, the exponent explodes and the density collapses and becomes exponentially small, which makes it very difficult for $\hat{\theta}$ to estimate $\tilde{\theta}$ precisely. Therefore we confine to $m = O(n)$ for overidentified models. It is also easy to show that

$$\frac{1}{m} \det\left(m^{d-1}\Sigma^{-1} - \frac{\sigma_0^2 m^2 \Sigma^{-1} ll' \Sigma^{-1}}{1 + \sigma_0^2 ml'\Sigma^{-1}l}\right) \to C > 0,$$

using the relation that $\det(I + uv') = 1 + u'v$. Furthermore, if the model is incorrectly specified, $\Delta \to \infty$. In this case $f_Y(0)$ is also exponentially small even when $m = n$. □

The general result mirrors this example.

LEMMA 2 Under Assumptions 1 to 4, $f_{Y_m^s}(0) / \sqrt{m}^{d-k} f_\infty(0) \xrightarrow{p} 1$, where

$$f_\infty(0) \equiv \det\left(W(\theta_0)\right)^{1/2} \exp\left(-\frac{m}{n}\frac{n}{2} \hat{g}\left(\tilde{\theta}\right)' W \hat{g}\left(\tilde{\theta}\right)\right).$$

This lemma includes several possibilities. In an exactly identified model where $d = k$, and $\hat{g}\left(\tilde{\theta}\right) = 0$, $f_{Y_m^s}(0) = O_P(1)$. In a correctly specified and overidentified model, $\hat{g}\left(\tilde{\theta}\right) = O_P\left(1/\sqrt{n}\right)$. In this case if $m/n = O(1)$, then $f_{Y_m^s}(0) = O_P\left(\sqrt{m}^{d-k}\right)$. If the overidentified model is misspecified, or if $m/n \to \infty$ in a correctly specified and overidentified model,

$f_{Y_m^s}(0)$ is exponentially small when the sample size n increases. In the misspecified case, $f_{Y_m^s}(0) = O_p(\exp(-nc))$ for some $c > 0$. In the correct but overidentified case $f_{Y_m^s}(0) = O_p(\exp(-m/n))$ when $m/n \to \infty$.

We saw again that in overidentified models, allowing for $m/n \to \infty$ causes implementation difficulties for $\hat{\theta}$ and $\hat{\eta}$ because of Y_m^s tends to be far from zero in finite sample, so that $f_Y(0)$ is very small. Therefore, while the asymptotic theory for the infeasible estimators $\bar{\theta}$ and $\bar{\eta}$ allows for $m \neq n$, it is natural in the implementation to choose $m = n$. In the following feasible implementation of $\hat{\theta}$ and $\hat{\eta}$, we first focus on $m = n$, and then comment on issues related to $m \neq n$. While allowing for $m \leq n$ saves on computation, inference is very conservative.

3.3 Asymptotic validity of estimators and posterior quantiles

We maintain a standard assumption for the kernel function.

ASSUMPTION 8 The kernel function satisfies (1) $\kappa(x) = h(|x|)$ where $h(\cdot)$ decreases monotonically on $(0, \infty)$; (2) $\int \kappa(x)\, dx = 1$; (3) $\int |x|^2 \kappa(x)\, dx < \infty$.

The following theorem is a consequence of the general result that $f(Y_n = 0)$ is bounded away from 0 and that $Var(\theta | Y_n = 0)$ is of the order of $O(1/n)$.

THEOREM 2 For the local constant kernel based estimator, in an exactly identified model, let $m = n$, under Assumptions 1-2 and 8, or Assumptions 1, 3-6 and 8, $\hat{\theta} \to_P \theta_0$ if $Sh^k \to \infty$ and $h \to 0$. Furthermore, if $Sh^k \to \infty$, $Sh^k \min\left(n, \frac{1}{h^2}\right) \to \infty$ and $\sqrt{n}h^2 \to 0$, then $\sqrt{n}\left(\hat{\theta} - \tilde{\theta}\right) \to_P 0$. These conditions are satisfied when $\sqrt{n}h = O(1)$ and $Sh^k \to \infty$. We may allow m/n to converge to ∞, but there is no first order efficiency gain. When $m/n \to 0$, a similar result that $\sqrt{m}\left(\hat{\theta} - \tilde{\theta}\right) \to_P 0$ holds if $\sqrt{m}h^2 = o(1)$, $Sh^k \to \infty$, and $Sh^k \min\left(m, \frac{1}{h^2}\right) \to \infty$, both of which are implied by $\sqrt{m}h = O(1)$. In an overidentified model, if $\sqrt{n}h \to \infty$, $\sqrt{n}h^2 \to 0$, $Sh^{k-2}/n \to \infty$, so that $\frac{1}{Sh^k}\left(\frac{1}{n} + h^2\right) << \frac{1}{n}$, then $\sqrt{n}\left(\hat{\theta} - \tilde{\theta}\right) \to_P 0$.

This result is mostly relevant for exactly identified models. When $m = n$, one only needs $Sh^k \to 0$ and $\sqrt{n}h = O(1)$ in order for $\hat{\theta}$ to be first order asymptotically equivalent to $\tilde{\theta}$.

The reason for this is that whenever $Sh^k \to \infty$, aside from the bias term, $\hat{\theta}$ is automatically $\sqrt{n}$ consistent for $E\left[\hat{\theta}\right]$. Hence interaction between the sample size n and the bandwidth is only relevant for the "bias" term. However, this bias term is different from the conventional bias term as it actually enters the "variance" estimation, and can not be improved using higher order kernels. The conventional bias term is of order $O\left(h^2\right)$ for a regular second order kernel, and $O\left(h^\gamma\right)$ for a higher γ–th order kernel. In order for the conventional bias term not to affect the asymptotic distribution, it only requires $\sqrt{n}h^2 \to 0$. But the bias term also contributes to the aggregate variance, so that the variance of $\sqrt{Sh^k}\left(\hat{\theta} - \bar{\theta}\right)$ is of the order of $1/n + h^2$. In order for the variance to be order $1/n$, we require $\sqrt{n}h = O\left(1\right)$. Consequently, the minimum condition on S so that $\sqrt{n}\left(\hat{\theta} - \bar{\theta}\right)$ is $S >> n^{k/2}$. This is not that different from the condition on the number of grids for a brute force grid search method to locate $\hat{\theta}$ up to a $O\left(1/\sqrt{n}\right)$ neighborhood. We conclude that kernel methods, regardless of the order of the kernel, can not improve much upon grid search in terms of reducing computational intensity.

With overidentification, we impose $\sqrt{n}h \to \infty$, this induces a stricter condition requiring a larger S: $Sh^{k-2}/n \to \infty$ instead of only $Sh^k \to \infty$. If we had chosen $\sqrt{n}h = O\left(1\right)$ for overidentified models, then the curse of dimensionality would have been in Sh^d instead of Sh^k, which is strictly worse than a brute force grid search.

As will be clear in the discussion below, the additional variance term can be reduced to the order of $\frac{1}{nh^4}$ instead of $\frac{1}{nh^2}$ when a local linear instead of local constant regression is used, where we define $\hat{\eta} = \hat{a}$, where

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \left(\sum_{s=1}^{S} \begin{pmatrix} 1 \\ y^s \end{pmatrix} \begin{pmatrix} 1 & y^s \end{pmatrix} \kappa\left(\frac{y^s}{h}\right)\right)^{-1} \left(\sum_{s=1}^{S} \begin{pmatrix} 1 \\ y^s \end{pmatrix} \eta^s \kappa\left(\frac{y^s}{h}\right)\right).$$

THEOREM 3 Using local linear regression, $\sqrt{n}\left(\hat{\eta} - \bar{\eta}\right) = o_P\left(1\right)$ when $Sh^k \to \infty$, $\sqrt{n}h \to \infty$ and $\sqrt{n}h^2 = o\left(1\right)$.

In particular, note that the extra local constant condition $\sqrt{n}h \to c < \infty$ is no longer needed with local linear regressions. Instead, we impose the additional assumption that $\sqrt{n}h \to \infty$ to simplify the proof and to allow for a smaller S. When $m \leq n$, it is possible to allow for smaller h but the proof needs to be revised and S also needs to be much larger.

But $m/n \to \infty$ will require $\sqrt{n}h \to \infty$. In this case, when h is too small, there might be too little data in a neighborhood of zero for the moment conditions to allow for accurate estimation of the parameters.

We also note that regardless of whether the model is overidentified, the curse of dimensionality is reflected in k, the number of parameters, and not in d, the dimension of the moment conditions.

Next, we prove the validity of the simulated posterior quantiles for frequentist inference. To allow for weaker conditions on the bandwidth parameter we focus on local linear regression.For this purpose, define $\hat{\eta}_\tau = \hat{a}$, as in

$$\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \rho_\tau \left(\eta^s - a - by^s\right) \kappa\left(\frac{y^s}{h}\right).$$

The following theorem holds.

THEOREM 4 Under Assumptions 1, 2, 3 and 8, if $nh^4 \to 0$, $\sqrt{n}h \to \infty$ and $Sh^k \to \infty$, which requires $S >> n^{k/4}$, $\hat{\eta}_\tau - \bar{\eta}_\tau = o_P\left(\frac{1}{\sqrt{n}}\right)$, so that posterior inference based on $\hat{\eta}_\tau$ is valid whenever it is valid for the infeasible $\bar{\eta}_\tau$.

The calculation in the above proof also suggests how to generalize the relevant conditions to high order local polynomials and higher order kernel functions. Using higher (pth order) polynomial and a regular kernel function, the required assumptions are weakened to $\sqrt{n}h^{p+1} \to 0$ and $Sh^k \to \infty$, which requires $S >> n^{\frac{k}{2(p+1)}}$. More generally, let p_1 be the order of the local polynomial used and let p_2 be the order of the kernel function. Then we have, given the data ($\tilde{\theta}$):

$$Var\left(\hat{\theta} - \tilde{\theta}\right) = \frac{1}{Sh^k}\left(\frac{1}{n} + h^{2(p_1+1)}\right), \qquad Bias\left(\hat{\theta} - \tilde{\theta}\right) = O\left(h^{\max(p_1+1,p_2)}\right).$$

Furthermore, $Var\left(\tilde{\theta} - \theta_0\right) = O\left(1/n\right)$. Therefore, the results of Theorems 2 and 4 hold under the alternative conditions that $\sqrt{n}h^{\max(p_1+1,p_2)} \to 0$, $Sh^k \to \infty$, and $\sqrt{n}h^{p_1+1} = O\left(1\right)$.

In summary, while both higher order polynomials and kernel methods reduce bias, higher order polynomials also improve on variance but kernel methods do not. A larger value of p_1 allows for a larger bandwidth h and a smaller number of simulations S.

That the curse of dimensionality is only reflected in k but not in d is due to the multicolinearity of moment conditions when $d > k$. We use Example 1 to illustrate the singularity of $f(Y)$ through a change of variable.

Example 1 continued For simplicity let $\Sigma = I$. Partition $Y = (Y_1, Y_2)$ for a scalar Y_1. Let $Y_2 = \ell Y_1 + \frac{\Delta}{\sqrt{n}} + \frac{w_2}{\sqrt{m}}$. Then

$$\begin{aligned} Y_1 =& \mu - \bar{X}_1 + \frac{\xi_1}{\sqrt{m}} \quad Y_2 = \mu - \bar{X}_2 + \frac{\xi_2}{\sqrt{m}}, \\ \Delta =& -\sqrt{n}\left(\bar{X}_2 - \bar{X}_1\right) = O_p(1) \quad w_2 = \xi_2 - \xi_1 = O_p(1). \end{aligned}$$

The implication of this for the kernel function is that

$$\kappa\left(\frac{Y_1}{h}, \frac{Y_2}{h}\right) = \kappa\left(\frac{Y_1}{h}, \frac{Y_1}{h} + \frac{\Delta}{\sqrt{n}h} + \frac{w_2}{\sqrt{m}h}\right).$$

If both $\sqrt{n}h \to \infty$ and $\sqrt{m}h \to \infty$, then $\frac{\Delta}{\sqrt{n}h} = o_p(1)$, $\frac{w_2}{\sqrt{m}h} = o_P(1)$, and essentially,

$$\kappa\left(\frac{Y_1}{h}, \frac{Y_2}{h}\right) \approx \kappa\left(\frac{Y_1}{h}, \ell\frac{Y_1}{h}\right) = \bar{\kappa}\left(\frac{Y_1}{h}\right)^d,$$

which resembles a one-dimensional kernel function.

The change of variables carries over to a more general setting. Partition again

$$Y = \left(Y_1 \in R^k, Y_2 \in R^{d-k}\right), \qquad g(\theta) = \left(\hat{g}_1(\theta) \in R^k, \hat{g}_2(\theta) \in R^k\right)$$

correspondingly, as well as $g(\theta) = (g_1(\theta), g_2(\theta))$. Consider now an (infeasible and only used in the proof construct) change of variable $y_2 = g_2\left(g_1^{-1}(y_1)\right) + \frac{\Delta}{\sqrt{n}}$.

Then, we can write $\Delta = O_p(1)$ since,

$$\begin{aligned} \Delta =& \sqrt{n}\left(\hat{g}_2(\theta) + \frac{\epsilon_2}{\sqrt{m}} - \left(g_2\left(g_1^{-1}\left(\hat{g}_1(\theta) + \frac{\epsilon_1}{\sqrt{m}}\right)\right)\right)\right) \\ =& \sqrt{n}\left(\hat{g}_2(\theta) - g_2(\theta)\right) + \sqrt{\frac{n}{m}}\epsilon_2 - \frac{\partial g_2(\theta)}{\partial\theta}\left(\frac{\partial g_1(\theta)}{\partial\theta}\right)^{-1}\sqrt{n}\left(\hat{g}_1(\theta) - g_1(\theta) + \frac{\epsilon_1}{\sqrt{m}}\right) + o_P(1). \end{aligned}$$

When $\sqrt{n}h \to \infty$, the original d dimensional kernel function becomes effectively k dimensional only, since

$$\kappa\left(\frac{y_1}{h}, \frac{y_2}{h}\right) = \kappa\left(\frac{y_1}{h}, \frac{g_2\left(g_1^{-1}(y_1)\right)}{h} + \frac{\Delta}{\sqrt{n}h}\right) \to \kappa\left(\frac{y_1}{h}, \frac{g_2\left(g_1^{-1}(y_1)\right)}{h}\right).$$

Intuitively, for $h >> 1/\sqrt{n}$, the bandwidth along the manifold of $y_2 = g_2\left(g_1^{-1}(y_1)\right)$ is $\frac{1}{\sqrt{n}}$ instead of h, and most effective observations fall within this narrow band around the manifold. Hence the effective number of observations is, $f_y(0) Sh^k \frac{1}{\sqrt{n}^{d-k}} = O_P\left(Sh^k\right)$ as $f_y(0) = O_P\left(\sqrt{n}^{d-k}\right)$.

If we define $c = \lim_{h\to 0} \frac{1}{h} g_2\left(g_1^{-1}(h)\right) = \nabla g_2(\theta_0)\left(\nabla g_1(\theta_0)\right)^{-1}$, then $g_2\left(g_1^{-1}(uh)\right) \sim cu$ and by a change of variable $y_1 = uh$,

$$\kappa\left(\frac{y_1}{h}, \frac{y_2}{h}\right) = \kappa\left(u, \frac{g_2\left(g_1^{-1}(uh)\right)}{h}\right) \to \kappa(u, cu).$$

For the nonparametric regression of θ on y, these calculations imply that the variance in local constant kernel methods is of order $\frac{1}{Sh^k}\left(\frac{1}{n} + h^2\right)$. In local linear regressions, however, regressors can be asymptotically collinear along a k dimensional manifold, with up to $1/\sqrt{n}$ local variation surrounding this manifold. Because of this, while the intercept term will converge at the fast $\frac{1}{\sqrt{Sh^k}}\frac{1}{\sqrt{n}}$ rate up to the bias adjustment term, the local linear coefficients will converge at the slower $\frac{1}{\sqrt{Sh^k}}$ rate. However, certain linear combinations of the coefficients converge at the faster rate of $\frac{1}{\sqrt{Sh^k}h}\frac{1}{\sqrt{n}}$. We illustrate this point in the normal analytic example 1 for $k = 1$, $d = 2$ with diffuse prior $\sigma_0 = \infty$.

Consider again the normal example when $m = n$: $\mu = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \epsilon$, where

$$\beta_0 = \left(l'\Sigma^{-1}l\right)^{-1} l'\Sigma^{-1}\bar{X}, \quad (\beta_1\ \beta_2) = \left(l'\Sigma^{-1}l\right)^{-1} l'\Sigma^{-1}, \quad \epsilon \sim N\left(0, \frac{1}{n}\left(l'\Sigma^{-1}l\right)^{-1}\right).$$

This can be written as

$$\begin{aligned}
\mu =& \beta_0 + Y_1(\beta_1 + \beta_2) + (Y_2 - Y_1)\beta_2 + \epsilon \\
\equiv& \beta_0 + Y_1\eta + \left(\bar{X}_1 - \bar{X}_2 + \frac{\xi_2}{\sqrt{n}} - \frac{\xi_1}{\sqrt{n}}\right)\beta_2 + \epsilon \\
=& \beta_0 + \left(\bar{X}_2 - \bar{X}_1\right)'\beta_2 + \left(\mu + \frac{\epsilon_1}{\sqrt{n}}\right)\eta + \epsilon_2\frac{\beta_2}{\sqrt{n}} + \epsilon.
\end{aligned}$$

Then for $\theta = \left(\beta_0 + \left(\bar{X}_2 - \bar{X}_1\right)'\beta_2, \beta_1 + \beta_2, \frac{\beta_2}{\sqrt{n}}\right)$ and its corresponding least squares estimate $\hat{\theta}$ based on the dependent variable $\mu_s, s = 1, \ldots, \bar{S}$ and regressors $Y_1 \equiv \mu + \frac{\epsilon_1}{\sqrt{n}}$ and $\sqrt{n}(Y_2 - Y_1) \equiv \epsilon_2$, where $\bar{S}$ is typically Sh^k, $\sqrt{\bar{S}}\left(\hat{\theta} - \theta\right)$ has a nondegenerate distribution. As $\bar{S} \to \infty$

$$\sqrt{n\bar{S}}\left(\hat{\theta} - \theta\right) \sim N\left(0, \sigma^2\Sigma_n^{-1}\right),$$

where

$$\Sigma_n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_u^2 + \frac{\sigma_1^2}{n} & \frac{\sigma_{12}}{\sqrt{n}} \\ 0 & \frac{\sigma_{12}}{\sqrt{n}} & \sigma_2^2 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{pmatrix}.$$

Asymptotically, $\hat{\beta}_1 + \hat{\beta}_2$ and $\hat{\beta}_2$ are independent. If $m/n \to \infty$, then

$$\epsilon \sim N\left(0, \frac{1}{m}\left(l'\Sigma^{-1}l\right)^{-1}\right),$$

and

$$Y_2 - Y_1 = \bar{X}_1 - \bar{X}_2 + \frac{\xi_2 - \xi_1}{\sqrt{m}}.$$

In this case, β_2 needs to be rescaled by $\sqrt{m}$ instead.

If $m = \infty$, Y_2 is a constant shift of Y_1 by $\bar{X}_2 - \bar{X}_1$, resulting in regressors being multicollinear. In this case, it appears sufficient to regress μ_s on either Y_{1s} or Y_{2s}. However, in this case the intercept term β_0 is not estimable within a scale of $O_p\left(1/\sqrt{n}\right)$, in the sense that $\hat{\beta}_0 = F'\hat{\beta}$ for $F = (1\ 0\ 0)'$ and $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ is not uniquely determined when $\hat{\beta}$ can have multiple solutions to the normal equation due to the collinearity of regressors.

To see this, note that the model can be transformed into one of the following two cases:

$$\mu = \left(\beta_0 + \left(\bar{X}_2 - \bar{X}_1\right)'\beta_2\right) + Y_1'\beta_2 \quad \text{or} \qquad \mu = \left(\beta_0 - \left(\bar{X}_2 - \bar{X}_1\right)'\beta_1\right) + Y_2'\beta_2.$$

Alternatively, we can use the arguments on pp 58, Amemiya (1985) that there is no A such that $F = X'A$. In fact, let $A = (a_1, \ldots, a_n)'$. Then to satisfy

$$(1\ 0\ 0)' = \left(\sum a_i \ \sum a_i Y_{1i} \ \left(\bar{X}_2 - \bar{X}_1\right)\sum a_i + \sum a_i Y_{1i}\right).$$

It is necessary that both $\sum a_i = 1$ and $\sum a_i = 0$. This forms a contradiction.

The collinear case is perhaps unlikely in nonlinear models. Nevertheless this suggests that in overidentified models, the limit of taking $m \to \infty$ rapidly can be different from setting $m = \infty$. It is obviously possible to regress only on Y_1 or Y_2, but in this case $\bar{X}_1$ and $\bar{X}_2$ might not have been combined optimally. In the proofs in the appendix, these calculations are demonstrated in the general case.

3.4 Local polynomial estimation

It is striaghtforward but tedious, to generalize the previous results on local linear mean and quantile regressions to formalize local polynomials combining the notations in Chaudhuri (1991) and the proof technique in Fan, Hu, and Truong (1994).

For $u = (u_1, \ldots, u_d)$, a d-dimensional vector of nonnegative integers, let $[u] = u_1 + \cdot + u_d$. Let A be the set of all d-dimensional vectors u such that $[u] \leq p$ and set $s(A) = \#(A)$. Let $\beta = (\beta_u)_{u \in A}$ be a vector of coefficients of dimension $s(A)$. Also let

$$y_s^A = \left(y_s^u = y_{s1}^{u_1} \ldots y_{sd}^{u_d}, u \in A\right)'.$$

Define the polynomial

$$P_n(\beta, y_s) = \sum_{u \in A} \beta_u y_s^u = \beta' y_s^A.$$

Then define the local polynomial mean estimator as $\hat{\theta} = \hat{\beta}_{[0]}$, the 0th element of $\hat{\beta}$, for

$$\hat{\beta} = \left(\sum_{s=1}^{S} y_A^s y_A^{s'} \kappa\left(\frac{y^s}{h}\right)\right)^{-1} \left(\sum_{s=1}^{S} y_s^A \theta^s \kappa\left(\frac{y^s}{h}\right)\right).$$

Similarly, we use local polynomial quantile regression to define $\hat{\eta}_\tau = \hat{\beta}_{[0]}$, for

$$\hat{\beta} \equiv \arg\min_{\beta} \sum_{s=1}^{S} \rho_\tau\left(\eta^s - \beta' y_s^A\right) \kappa\left(\frac{y_s}{h}\right).$$

THEOREM 5 Under Assumptions 1-2 and 8, or Assumptions 1, 3-6 and 8, if $nh^{2(p+1)} \to 0$, $\sqrt{n}h \to \infty$, $Sh^k \to \infty$ (which requires that $S >> n^{\frac{k}{2(p+1)}}$), $\hat{\theta} - \bar{\theta} = o_p(1/\sqrt{n})$. Furthermore, $\hat{\eta}_\tau - \bar{\eta}_\tau = o_p(1/\sqrt{n})$, so that posterior inference based on $\hat{\eta}_\tau$ is valid whenever it is valid for the infeasible $\bar{\eta}_\tau$.

The proof is analogous to the local linear case. The appendix provides the key steps in the proof for completeness.

3.5 Nearest neighborhood implementation

One possible method to choose the window width parameter h is using nearest neighborhood of zero of the moment conditions. Instead of choosing h, the researcher picks a nearest

neighbor number κ_n that is dependent on the sample size. The simulated draws $Y_m^s, s = 1, \ldots, S$ are sorted according to a suitable norm $|Y_m^s|, s = 1, \ldots, S$, that can be for example the usual Euclidean norm. Heuristically, one may also sort $s = 1, \ldots, S$ based on the GMM objective function $\hat{g}(\theta^s)' \hat{W} \hat{g}(\theta^s)$. Collect the k_n elements of $s = 1, \ldots, S$ such that $|Y_m^s|$ or $\hat{g}(\theta^s)' \hat{W} \hat{g}(\theta^s)$ are the closest to zero in ascending order. Then the bandwidth parameter h can be chosen to be the distance of the κ_nth element of this set to zero: $h = |Y_m^{\kappa_n}$ or $h = \hat{g}(\theta^{\kappa_n})' \hat{W} \hat{g}(\theta^{\kappa_n})$. It is possible to show that $\kappa_n = O\left(nh^k\right)$. Therefore, for example, if $h = o(n^{-\frac{1}{2(p+1)}})$, then $\kappa_n = o\left(n^{1-\frac{k}{2(p+1)}}\right)$. Unlike the kernel method where the estimation window might be empty, the nearest neighborhood method will always produce a numerical estimate even when the model is misspecified.

4 Monte Carlo Simulation

In this section we report a Monte Carlo simulation based on the quantile instrumental variable model of Chernozhukov and Hansen (2005), which is not separable between the parameters and the data. In the model

$$y_i = x_i' \beta + \epsilon_i,$$

where $Q_\tau(\epsilon_i | z_i) = 0$. We consider the following data generating process:

$$\epsilon_i = \exp\left(\left(z_i' \alpha\right)^2 v_i\right) - 1,$$

where v_i is such that $Q_\tau(v_i | z_i) = 0$. In particular, we choose $x_i = (1, \tilde{x}_i)$ where $\tilde{x}_i = \xi_{1i} + \xi_{2i}$, $z_i = (1, \tilde{z}_i^1, \tilde{z}_i^2)$ where

$$\tilde{z}_i^1 = \xi_{2i} + \xi_{3i}, \qquad \tilde{z}_i^2 = \bar{\xi}_{i1} + \xi_{4i}$$

such that $\bar{\xi}_{i1}$ is a subset of ξ_{i1}, and $v_i \sim N(0, 1)$. In the above $\xi_{1i}, \xi_{2i}, \xi_{3i}, \xi_{4i}$ are independent vectors of normal variates with variances $\Omega_1, \ldots, \Omega_4$. Input parameters for the simulation include $\alpha, \beta, \Omega_1, \ldots, \Omega_4$. The parameter of interest is β, whose estimation is based on the moment condition

$$\hat{g}(\beta) = \frac{1}{n} \sum_{i=1}^{n} z_i \left(\tau - 1\left(y_i \leq x_i' \beta\right)\right).$$

The optimal weight matrix does not depend on parameters and is the inverse of $\frac{1}{n}\sum_{i=1}^{n} z_i z_i'$. We choose $\Omega_1 = \ldots = \Omega_4 = I$, and $\beta = (1,1)$. To implement the estimators we use a Gaussian kernel, and choose $h = \text{factor1} \times n^{-1/4}$ and $S = \text{factor2} \times h^{-d}$. We vary the sample size, the auxiliary parameters, and the rules for choosing the bandwidth and the number of simulations as follows: $n = (200, 400, 800)$; $\alpha = (1/\text{varscale}, 1/\text{varscale})$; varscale $= (5, 7)$; factor1 $= (0.05, 0.2, 0.4)$ and factor2 $= (40, 80, 160)$. The total number of outerloop Monte Carlo simulations is 100.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

Tables 1-4 report the simulation results for a selection of parameter configurations. These results show that local linear mean and median regressions outperform local constant mean and median regressions in terms of both bias and variance, and hence in mean square errors. The rows for "prior mean" are based on two stage least-square regression, which has larger mean square errors than both local constant and local linear regressions. Coverage probabilities for nominal 90% confidence intervals based on posterior quantiles are still far from ideal, but local linear quantile regression based intervals do seem to be closer to the nominal rate.

5 Empirical Illustration

Empirically, often times n, d and k can all be large. For example, in Angrist and Krueger (1991), $n = 329509$, $k = 11$, $d = 40$. The requirements on S and h are mild when the sample size n is moderate, but can be strigent when n is large and when k is not small. For large n and k, the computer memory requirements for large S can be excessive. This will require either huge memory or large parallel computing facilities. Alternatively, when memory is small and when S is large, it is also easy to compute the local constant and local linear

least squares estimators by sequentially updating the summation over S or by breaking this summation into sequences. It is an open question of whether it is possible to compute the local quantile estimators by sequentially updating or by summing over segments of summary statistics, in order to relieve the memory burden for large S.

Alternatively, one may wish to use a large value of h and a smaller number of simulations S that do not satisfy the above conditions. In this case, the sampling error will be dominated by the simulation error. One possibility is to develop an approximate distribution for $\hat{\theta} - \theta_0$ that is dominated by simulation variance. However, this is not possible when $p_2 = p_1 + 1 = p$, because when $Sh^k \to \infty$, the variance term $\frac{1}{Sh^k}\left(\frac{1}{n} + h^{2p}\right)$ is always smaller than the bias term h^{2p}. However, when $p_2 > p_1 + 1$, one may choose S and h in such a way that $\frac{1}{Sh^k}\left(\frac{1}{n} + h^{2(p_1+1)}\right) >> h^{2p_2}$ even if $h^{2(p_1+1)} >> \frac{1}{n}$ and $h^{2p_2} >> \frac{1}{n}$. Then one may find a way to estimate the variance term induced by the simulation error and use it to form inference statements. This still requires S and h be chosen in such a way that $h << n^{-1/(2p_2)}$.

Angrist and Krueger (1991) study the empirical model:

$$Y = \alpha + \beta D + \eta X$$

In the above, Y is log wage, D is education, and X are covariates including year of birth dummies. The instruments Z include birth quarters interacted with year of birth dummies.

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

Due to memory limitation, we choose $S = 7000$ and $h = S^{-\frac{1}{k+\delta}}$ for a small δ. The prior variance is chosen to be a multiple of the 2SLS variance. Because of the large sample size, the 2SLS variance is very small. The results are only visibly different from 2SLS when the prior variance is a very large multiple of the 2SLS variance. For local constant and linear mean regression, it is possible to bypass memory limitation by serialization. The memory requirement for local constant quantiles can also be reduced to a lesser extent. However, we are not yet able to serialize quantile regression. The next set of tables report these results for $S = 100,000$.

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

6 Conclusion

In this paper we have proposed a method of implementing quasi-Bayesian estimators for GMM models allowing for nonlinearity and nonseparability that is motivated also by indirect inference. This method combines simulation with nonparametric regression, and complements and extends Chernozhukov and Hong (2003) and Creel and Kristensen (2011). We consider both kernel and local linear (polynomials) methods, and demonstrate the asymptotic validity of posterior quantile based inference. In future work, we intend to formalize local polynomial methods and to study the combination with sieve and bootstrap methods.

A Proof of Theorem 1

Take $\theta^* = \theta^0$, then Assumptions 1 and 2 imply, by the arguments in Chernozhukov and Hong (2003), that $\bar{\theta} - \theta_0 = o_P(1)$, and that

$$\tilde{\theta} - \theta_0 = J^{-1}\Delta_n + o_P\left(1/\sqrt{n}\right).$$

Now for $\theta^* = \tilde{\theta}$, by Assumption 2, for $\theta - \hat{\theta} = o_P(1)$,

$$m\left(\hat{Q}_n(\theta) - \hat{Q}_n\left(\hat{\theta}\right)\right) = -\frac{1}{2}\sqrt{m}\left(\theta - \hat{\theta}\right)' J\sqrt{m}\left(\theta - \hat{\theta}\right) + o_P\left(m|\theta - \hat{\theta}|^2\right).$$

Define $V = \sqrt{m}\left(\theta - \hat{\theta}\right)$. The same arguments in Chernozhukov and Hong (2003) show that

$$\int v^\alpha |f_V(v|Y_m = 0) - p_\infty(v)|dv = o_P(1), \qquad \text{for} \quad p_\infty(v) = \sqrt{\frac{\det(J)}{\left(\sqrt{2\pi}\right)^k}} exp\left(-\frac{1}{2}v'Jv\right)$$

and for all $\alpha \geq 0$. Consequently,

$$\sqrt{m}\left(\theta - \tilde{\theta}\right) \to_d N\left(0, J^{-1}\right) \quad \text{and} \quad \sqrt{m}\left(\bar{\theta} - \tilde{\theta}\right) = o_P(1).$$

Then by the Delta method, for $\tilde{\eta} = \eta\left(\tilde{\theta}\right)$,

$$\sqrt{m}\left(\eta\left(\theta\right) - \tilde{\eta}\right) \to_d N\left(0, \sigma_\eta^2\right),$$

where $\sigma_\eta^2 = \frac{\partial}{\partial\theta}\eta_0' J^{-1} \frac{\partial}{\partial\theta}\eta_0$.

Then $\sqrt{m}\left(\bar{\eta}_{1/2} - \tilde{\eta}\right) \to_P 0$, $\sqrt{m}\left(\bar{\eta}_\tau - \bar{\eta}_{1/2}\right) \to_P \Phi^{-1}\left(\tau\right)\sigma_\eta$. Now consider separately $m \geq n$ and $m/n \to 0$. In the first case,

$$\sqrt{n}\left(\bar{\eta}_{1/2} - \eta_0\right) = \sqrt{n}\left(\tilde{\eta} - \eta_0\right) + o_P\left(1\right) \to_d N\left(0, \sigma_\eta^2\right).$$

(Recall the information matrix equality assumption that $\Omega = J$.) Then

$$\begin{aligned}
&P\left(\eta_0 \in \left(\bar{\eta}_{1/2} - \sqrt{\frac{m}{n}}\left(\bar{\eta}_{1-\tau/2} - \bar{\eta}_{1/2}\right), \bar{\eta}_{1/2} - \sqrt{\frac{m}{n}}\left(\bar{\eta}_{\tau/2} - \bar{\eta}_{1/2}\right)\right)\right) \\
&= P\left(-\sqrt{m}\left(\bar{\eta}_{1-\tau/2} - \bar{\eta}_{1/2}\right) \leq \sqrt{n}\left(\tilde{\eta} - \eta_0\right) + o_P\left(1\right) \leq -\sqrt{m}\left(\bar{\eta}_{\tau/2} - \bar{\eta}_{1/2}\right)\right) \to 1 - \tau.
\end{aligned}$$

In the second case where $m/n \to 0$,

$$\sqrt{m}\left(\bar{\eta}_{1/2} - \eta_0\right) = \sqrt{m}\left(\bar{\eta}_{1/2} - \tilde{\eta}\right) + o_P\left(1\right) \to_P 0.$$

Then

$$\begin{aligned}
&P\left(\eta_0 \in \left(\bar{\eta}_{1/2} - \left(\bar{\eta}_{1-\tau/2} - \bar{\eta}_{1/2}\right), \bar{\eta}_{1/2} - \left(\bar{\eta}_{\tau/2} - \bar{\eta}_{1/2}\right)\right)\right) \\
&= P\left(-\sqrt{m}\left(\bar{\eta}_{1-\tau/2} - \bar{\eta}_{1/2}\right) \leq \sqrt{m}\left(\bar{\eta}_{1/2} - \eta_0\right) \leq -\sqrt{m}\left(\bar{\eta}_{\tau/2} - \bar{\eta}_{1/2}\right)\right) \\
&= P\left(-\sqrt{m}\left(\bar{\eta}_{1-\tau/2} - \bar{\eta}_{1/2}\right) \leq o_P\left(1\right) \leq -\sqrt{m}\left(\bar{\eta}_{\tau/2} - \bar{\eta}_{1/2}\right)\right) \to 1.
\end{aligned}$$

Therefore, when $m \geq n$, posterior quantile inference is asymptotically exact when the model is correctly specified, or when the optimal weighting matrix is used. On the other hand, when $m/n \to 0$, posterior quantile inference is asymptotically conservative. In this sense, regardless of the size of the simulated sample m, posterior quantile inference is valid whenever the model is correctly specified.

B Proof of Lemma 2

First note that for each v, $f_{\sqrt{m}\left(\theta - \tilde{\theta}\right)}\left(v|Y_m = t\right) = \left(\sqrt{m}\right)^{-k} f_\theta\left(\tilde{\theta} + \frac{v}{\sqrt{m}}|Y_m = t\right)$ and

$$f_\theta\left(\tilde{\theta} + \frac{v}{\sqrt{m}}|Y_m = t\right) = \frac{\pi_\theta\left(\tilde{\theta} + \frac{v}{\sqrt{m}}\right) f\left(Y_m = t|\tilde{\theta} + \frac{v}{\sqrt{m}}\right)}{f\left(Y_m = t\right)}.$$

Therefore one can write

$$f\left(Y_m = t\right) = \sqrt{m}^{-k} \frac{\pi_\theta\left(\tilde{\theta} + \frac{v}{\sqrt{m}}\right) f\left(Y_m = t | \theta = \tilde{\theta} + \frac{v}{\sqrt{m}}\right)}{f_{\sqrt{m}\left(\theta - \tilde{\theta}\right)}\left(v | Y_m = t\right)}.$$

Since both $f_{\sqrt{m}\left(\theta-\tilde{\theta}\right)}\left(v|Y_m = t\right) = O\left(1\right)$ and $\pi_\theta\left(\tilde{\theta} + \frac{v}{\sqrt{m}}\right) = O\left(1\right)$,

$$f\left(Y_m = t\right) \approx C\sqrt{m}^{-k} f\left(Y_m = t | \theta = \tilde{\theta} + \frac{v}{\sqrt{m}}\right).$$

Consider for example the case of $t = v = 0$, then

$$f\left(Y_m = 0 | \theta = \tilde{\theta}\right) \propto m^{d/2} \det\left(W\right)^{1/2} \exp\left(-\frac{m}{n}\frac{n}{2}\hat{g}\left(\tilde{\theta}\right)' W \hat{g}\left(\tilde{\theta}\right)\right).$$

Since $\frac{n}{2}\hat{g}\left(\tilde{\theta}\right)' W \hat{g}\left(\tilde{\theta}\right) = O_P\left(1\right)$ when the model is correctly specified, when $m = O(1)$, $f\left(Y_m = 0\right) \approx O\left(m^{\frac{d-k}{2}}\right)$. The exact identification case is a special case of this when $d = k$. The density of $f\left(Y_m = 0\right)$ will collapse exponentially if either the model is misspecified, or $m/n \to \infty$. More precisely, for $m/n \to c < \infty$ and $\sqrt{n}\hat{g}\left(\tilde{\theta}\right) \to_d Z$ for a random variable Z,

$$\frac{1}{\sqrt{m}^{d-k}} f\left(Y_m = 0\right) - \frac{\pi\left(\theta_0\right) \det\left(W\right)^{1/2} e^{-\frac{1}{2}cZ'WZ}}{\rho_\infty\left(0\right)} \to_P 0.$$

C Proof of Theorem 2

First we consider the expression of the asymptotic bias, which is related to the derivative of the moment condition with respect to local misspecification. For each t, define

$$\tilde{\theta}\left(t\right) = \arg\max_\theta \hat{Q}\left(\theta, t\right) \equiv \left(\hat{g}\left(\theta\right) - t\right)' \hat{W} \left(\hat{g}\left(\theta\right) - t\right).$$

Also define its population analog

$$\theta\left(t\right) = \arg\max_\theta Q\left(\theta, t\right) \equiv \left(g\left(\theta\right) - t\right)' W \left(g\left(\theta\right) - t\right).$$

Note that $\bar{\theta}\left(t\right) = E\left(\theta | Y_m = t\right)$. Then

$$\bar{\theta}\left(t\right) - \tilde{\theta}\left(t\right) = o_P\left(1\sqrt{m}\right) \quad \text{and} \quad \tilde{\theta}\left(t\right) - \theta\left(t\right) = O_P\left(1/\sqrt{n}\right),$$

where some extra derivations can show the uniformity of these convergence results in t. Then $\frac{\partial^2 \bar{\theta}(t)}{\partial t^2}$ converges to $\frac{\partial^2 \theta(t)}{\partial t^2}$. Note that $\theta(t)$ is defined by the solution to, for $G(\theta)$ being the Jacobian of the population moment condition $g(\theta)$,

$$G(\theta)' W (g(\theta) - t) = 0.$$

Totally differentiating this equality shows that, for $H_k(\theta) = \frac{\partial G(\theta)}{\partial \theta_k}$,

$$\frac{\partial \theta(t)}{\partial t} = \left(G(\theta_t)' W G(\theta_t) + \sum_k H_k(\theta_t)' W (g(\theta_t) - t) \right)^{-1} G(\theta_t)' W.$$

This can be further differentiated to give an analytic expression for $\frac{\partial^2 \theta(t)}{\partial t^2}$. In general, $g(\theta_t) - t = 0$ only when $t = 0$. When $t \neq 0$, the moment conditions $g(\theta_t) = t$ can be misspecified, as considered in Hall and Inoue (2003).

For $\bar{\theta}(t) = E(\theta | Y_m = t)$, we require an approximate Taylor expansion of $\bar{\theta}(t) = E(\theta | Y_m = t)$ in relation to $\theta(t)$. We need, for $t \to 0$,

$$\bar{\theta}(t) = \bar{\theta}(0) + a(0)' t + \frac{1}{2} t' b(0) t + o_P\left(\frac{1}{\sqrt{n}}\right) + o\left(t^2\right).$$

This will follow from showing that

$$\bar{\theta}(t) = \bar{\theta}(0) + \theta(t) - \theta(0) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

To see this, note that under Assumptions 1-2, or Assumptions 1, 3-6, the arguments in Chernozhukov and Hong (2003) can be generalized to show that uniformly in t close to and including zero,

$$\sqrt{n}\left(\bar{\theta}(t) - \tilde{\theta}(t)\right) = o_P(1), \quad \sqrt{n}\left(\tilde{\theta}(t) - \theta(t)\right) = O_P(1).$$

These uniform rate results are shown in the technical addendum. Therefore in the following we will focus on $\tilde{\theta}(t)$.

Next it remains to show uniform stochastic equicontinuity, namely that for all $t \to 0$,

$$\sqrt{n}\left(\tilde{\theta}(t) - \theta(t) - \left(\tilde{\theta}(0) - \theta(0)\right)\right) = o_P(1). \tag{6}$$

This in turn follows from the following calculation. Denote $G_t = G(\theta_t)$, it is shown in the technical addendum that under the stated assumptions,

$$\begin{aligned}\sqrt{n}\left(\tilde{\theta}(t) - \theta(t)\right) = &-\left(G_t' W_t G_t + H_t' W_t (g(\theta_t) - t)\right)^{-1} \\ &\sqrt{n}\left[G_t W_t\left(\hat{g}(\theta_t) - g(\theta_t)\right)\right] + o_P(1)\end{aligned} \tag{7}$$

and

$$\sqrt{n}\left(\tilde{\theta}(0) - \theta(0)\right) = -\left(G'WG\right)^{-1}\sqrt{n}G'W\left(\hat{g}(\theta_0) - g(\theta_0)\right) + o_P(1)$$

As $t \to 0$, $g(\theta_t) \to g(\theta_0) = 0$, $G_t' W_t G_t + H_t' W_t (g(\theta_t) - t) \to G'WG$. In addition,

$$\sqrt{n}\left(\hat{g}(\theta_t) - g(\theta_t) - \hat{g}(\theta_0)\right) = o_P(1).$$

Recall that $g(\theta_0) = 0$ under correct specification. Hence we can write

$$\begin{aligned}&G_t W_t\left(\hat{g}(\theta_t) - g(\theta_t)\right) - G'W\hat{g}(\theta_0) = G_t W_t\left(\hat{g}(\theta_t) - g(\theta_t) - \hat{g}(\theta_0)\right) + \left(G_t W_t - G'W\right)\hat{g}(\theta_0) \\ = \quad & O(1)\, o_P\left(\frac{1}{\sqrt{n}}\right) + o(1)\, O_P\left(\frac{1}{\sqrt{n}}\right) = o_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Hence we conclude that (6) holds under correct specification when $g(\theta_0) = 0$. [1]

Note that (6) does not hold under misspecification when $g(\theta_0) \neq 0$. To illustrate consider smooth models. In this case

$$\sqrt{n}\left(\tilde{\theta}(0) - \theta(0)\right) = -\left(G_0' W_0 G_0 + H_0' W_0 g(\theta_0)\right)^{-1}\sqrt{n}\left(\hat{G}_0' \hat{W}_0 \hat{g}(\theta_0) - G_0' W_0 g(\theta_0)\right) + o_P(1).$$

While there is still

$$\sqrt{n}\left(\hat{g}(\theta_t) - g(\theta_t) - \hat{g}(\theta_0) + g(\theta_0)\right) = o_P(1),$$

[1] To illustrate (7), consider smooth models. Let $\hat{G}_t = \hat{G}(\theta_t)$. Note that

$$\begin{aligned}\sqrt{n}\left(\tilde{\theta}(t) - \theta(t)\right) = &-\left(G_t' W_t G_t + H_t' W_t (g(\theta_t) - t)\right)^{-1} \\ &\sqrt{n}\left[\hat{G}_t \hat{W}_t\left(\hat{g}(\theta_t) - t\right) - G_t W_t\left(g(\theta_t) - t\right)\right] + o_P(1) \\ = &-\left(G_t' W_t G_t + H_t' W_t (g(\theta_t) - t)\right)^{-1} \\ &\sqrt{n}\left[\left(\hat{G}_t \hat{W}_t - G_t W_t\right)\left(g(\theta_t) - t\right) - G_t W_t\left(\hat{g}(\theta_t) - g(\theta_t)\right)\right] + o_P(1).\end{aligned}$$

But since $g(\theta_t) - t = o(1)$, $\left(\hat{G}_t \hat{W}_t - G_t W_t\right)\left(g(\theta_t) - t\right) = o_p(1/\sqrt{n})$. The technical addendum shows that (7) also holds for certain nonsmooth models under suitable conditions.

it no longer translates into (6) because

$$\begin{aligned}
&\hat{G}_t\hat{W}_t\left(\hat{g}\left(\theta_t\right)-t\right)-G_tW_t\left(g\left(\theta_t\right)-t\right)-\left(\hat{G}_0\hat{W}_0\hat{g}\left(\theta_0\right)-G_0W_0g\left(\theta_0\right)\right)\\
&=\left(\hat{G}_t\hat{W}_t-G_tW_t\right)\left(g\left(\theta_t\right)-t\right)+\left(\hat{G}_t\hat{W}_t-\hat{G}_0\hat{W}_0\right)\left(\hat{g}\left(\theta_t\right)-g\left(\theta_t\right)\right)\\
&\qquad+\hat{G}_0\hat{W}_0\left(\hat{g}\left(\theta_t\right)-g\left(\theta_t\right)-\hat{g}\left(\theta_0\right)+g\left(\theta_0\right)\right)+\left(\hat{G}_0\hat{W}_0-G_0W_0\right)g\left(\theta_0\right).
\end{aligned}$$

While the other terms are all $o_P\left(\frac{1}{\sqrt{n}}\right)$, the last term $\left(\hat{G}_0\hat{W}_0-G_0W_0\right)g\left(\theta_0\right)$ is only $O_P\left(\frac{1}{\sqrt{n}}\right)$ when $g\left(\theta_0\right)\neq 0$.

In the next we consider the exact identification case and the overidentification case separately.

Exact identification $d=k$. Consider first

$$\hat{\theta}-\bar{\theta}=\frac{A_1+A_2}{A_3}$$

where $A_3=\hat{f}\left(Y_m=0\right)=\frac{1}{Sh^k}\sum_{s=1}^{S}\kappa\left(Y_m^s/h\right)$,

$$A_1=\frac{1}{Sh^k}\sum_{s=1}^{S}\left(\theta^s-\bar{\theta}\right)\kappa\left(Y_m^s/h\right)-E\left(\left(\theta^s-\bar{\theta}\right)\frac{1}{h^k}\kappa\left(Y_m^s/h\right)\right),$$

and $A_2=E\left(\left(\theta^s-\bar{\theta}\right)\frac{1}{h^k}\kappa\left(Y_m^s/h\right)\right)$. Then

$$A_3\to_P\frac{\pi\left(\theta_0\right)\det\left(W\right)^{1/2}e^{-\frac{1}{2}c\hat{g}\left(\tilde{\theta}\right)'W\hat{g}\left(\tilde{\theta}\right)}}{\rho_\infty\left(0\right)}.$$

Also $A_2/\left(h^2\right)-C\to_P 0$ for $C=O_P\left(1\right)$. In particular

$$C=\int u^2\kappa\left(u\right)du\left(\frac{1}{2}\bar{\theta}''\left(t\right)h\left(t\right)+\bar{\theta}'\left(t\right)h'\left(t\right)\right),$$

where

$$h\left(t\right)=\frac{\pi\left(\theta_0\right)\det\left(W\right)^{1/2}e^{-\frac{1}{2}c\left(\hat{g}\left(\tilde{\theta}\right)-t\right)'W\left(\hat{g}\left(\tilde{\theta}\right)-t\right)}}{\rho_\infty\left(t\right)}.$$

In fact we note that as

$$\theta|Y_s=t\sim N\left(\bar{\theta}+J^{-1}GWt,J^{-1}/m\right),$$

because $f(\theta|Y_s = t)$ is proportional to

$$\exp\left(-\frac{m}{2}\left(\hat{g}\left(\tilde{\theta}\right)-t\right)'W\left(\hat{g}\left(\tilde{\theta}\right)-t\right)+\frac{m}{2}\hat{g}\left(\tilde{\theta}\right)'W\hat{g}\left(\tilde{\theta}\right)\right),$$

$\theta(t)$ approaches a linear function in t: $\theta(t) \to \tilde{\theta} + J^{-1}GWt$. This characterizes $\theta'(t)$.

Consider now the variance of A_1:

$$\begin{aligned}
&Var(A_1) = \frac{1}{Sh^{2k}}Var\left(\left(\theta^s-\bar{\theta}\right)\kappa\left(Y_m^s/h\right)\right)\\
=&\frac{1}{Sh^{2k}}\left(E_{Y_m}Var\left(\theta^s|Y_m^s\right)\kappa^2\left(Y_m^s/h\right)+Var_{Y_m}\left(E\left(\theta|Y_m^s\right)-\bar{\theta}\right)\kappa\left(Y_m^s/h\right)\right)\\
\leq&\frac{1}{Sh^{2k}}\left(E_{Y_m}Var\left(\theta^s|Y_m^s\right)\kappa^2\left(Y_m^s/h\right)+E_{Y_m}\left(E\left(\theta|Y_m^s\right)-\bar{\theta}\right)^2\kappa^2\left(Y_m^s/h\right)\right).
\end{aligned}$$

In the usual situation, the first term in parenthesis is $O(1)$ while the second term is $o(1)$. However, in this case both are $o(1)$. In particular, $Var(\theta^s|Y_m^s) = O(1/m)$ and $\left(E\left(\theta|Y_m^s = uh\right)-\bar{\theta}\right) = O(h)$. Hence, under the condition that $mh^2 = O(1)$, which has been assumed, for some $C > 0$,

$$Var\left(\sqrt{Sh^k}\left(\sqrt{m}+\frac{1}{h}\right)A_1\right) - J^{-1}h(0)\int\kappa^2(u)\,du + C \to_P 0.$$

Hence, $A_2/(h^2A_3) - C/h(0) \to_P 0$, and

$$\sqrt{Sh^k\left(m+\frac{1}{h^2}\right)}\frac{A_1}{A_3} \to_d N\left(0, \frac{J^{-1}}{h(0)}\right).$$

Hence $A_1/A_3 = O_P\left(\sqrt{Sh^k\left(m+\frac{1}{h^2}\right)}^{-1}\right)$ and $A_2/A_3 = O_P(h^2)$. The stated result follows.

Over identification $d > k$, $m = n$. Consider the following transformation:

$$y_2 = g_2\left(g_1^{-1}(y_1)\right) + \frac{\Delta}{\sqrt{n}} = cy_1 + \left(g_2\left(g_1^{-1}(y_1)\right) - cy_1 + \frac{\Delta}{\sqrt{n}}\right).$$

and the following change of variable $w = (w_1, w_2)$, $w_1 = y_1$,

$$w_2 = \sqrt{n}\left(g_2\left(g_1^{-1}(y_1)\right) - cy_1 + \frac{\Delta}{\sqrt{n}}\right). \tag{8}$$

Note that

$$\kappa\left(\frac{y}{h}\right) = \kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right).$$

Standard arguments can be used to show that $A_3 = (1 + o_P(1)) EA_3$, where, by changing $w_1 = uh$,

$$\begin{aligned} EA_3 =& E\frac{1}{h^k}\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) = \int\int \frac{1}{h^k}\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) dF(w_1, w_2) \\ =& \int\int \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) f(uh, w_2)\, dudw_2 \to f_{w_1}(0) \int \kappa(u, cu)\, du. \end{aligned}$$

Next consider

$$\begin{aligned} EA_2 =& E\left(\theta\left(w_1, cw_1 + \frac{w_2}{\sqrt{n}}\right) - \theta(0)\right)\frac{1}{h^k}\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) \\ =& EA_2^1 + EA_2^2 + EA_2^3. \end{aligned}$$

In the above, $EA_2^3 = o(h^2)$, where

$$EA_2^3 = \int\int o\left(w_1^2 + \frac{w_2^2}{n}\right)\frac{1}{h^k}\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) f(w_1, w_2)\, dw_1 dw_2.$$

$EA_2^2/h^2 \to C_1$ where

$$EA_2^2 = \frac{1}{2}\int \begin{pmatrix} w_1 \\ cw_1 + \frac{w_2}{\sqrt{n}} \end{pmatrix}' \theta''(0) \begin{pmatrix} w_1 \\ cw_1 + \frac{w_2}{\sqrt{n}} \end{pmatrix} \frac{1}{h^k}\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) dF(w_1, w_2).$$

Next consider EA_2^1/h, and change $w_1/h = u$

$$\begin{aligned} & \theta'(0)\frac{1}{h}\int\int \begin{pmatrix} w_1 \\ cw_1 + \frac{w_2}{\sqrt{n}} \end{pmatrix} \frac{1}{h^k}\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) f(w_1, w_2)\, dw_1 dw_2 \\ =& \int\int \theta'(0) \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix} \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) f(uh, w_2)\, dudw_2 \\ =& \int\int \theta'(0) \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix} \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) \left[f(0, w_2) + f^{(1)}(0^*, w_2)\, uh\right] dudw_2 \end{aligned}$$

for mean value 0^* between 0 and uh. Furthermore, by definition (8), w_2 is asymptotically normal. Therefore, for a density $\bar{f}(0, w_2)$ that is symmetric in w_2 around zero, $f(0, w_2) = \bar{f}(0, w_2) + o_P(1)$. In fact, an Edgeworth type argument can be used to show that $f(0, w_2) = \bar{f}(0, w_2) + o_P\left(\frac{1}{\sqrt{n}}\right)$. For symmetric $\kappa(\cdot, \cdot)$ and $\bar{f}(0, w_2)$,

$$\int\int \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix} \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) \bar{f}(0, w_2)\, dudw_2 = 0.$$

Also

$$\int\int \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix} \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) \left(f^{(1)}(0^*, w_2)\, u\right) du dw_2$$
$$\to \int\int \begin{pmatrix} u^2 \\ cu^2 \end{pmatrix} \kappa(u, cu)\, f(0, w_2)\, du dw_2 \equiv C_2.$$

Hence we conclude that

$$EA_2 = h^2(C_1 + C_2) + o_P(h) \qquad \text{or} \qquad EA_2 = h^2(C_1 + C_2) + O_P\left(\frac{h}{\sqrt{n}}\right).$$

Next recall that

$$\begin{aligned} Var(A_1) &= \frac{1}{Sh^{2k}} Var\left(\left(\theta^s - \bar{\theta}\right)\kappa\left(Y_m^s/h\right)\right) \\ \leq &\frac{1}{Sh^{2k}}\left(E_{Y_m} Var\left(\theta^s | Y_m^s\right)\kappa^2\left(Y_m^s/h\right) + E_{Y_m}\left(E\left(\theta | Y_m^s\right) - \bar{\theta}\right)^2 \kappa^2\left(Y_m^s/h\right)\right). \end{aligned}$$

Note that

$$\frac{1}{h^k} E_{Y_m} Var\left(\theta^s | Y_m^s\right)\kappa^2\left(Y_m^s/h\right) = O\left(\frac{1}{m}\right) f_{w_1}(0) \int \kappa^2(u, cu)\, du.$$

Furthermore, simple derivation implies that

$$\frac{1}{h^{2+k}} E_{Y_m}\left(E\left(\theta | Y_m^s\right) - \bar{\theta}\right)^2 \kappa^2\left(Y_m^s/h\right) \to f_{w_1}(0) \int \left(\theta'(0) \begin{pmatrix} u \\ cu \end{pmatrix}\right)^2 \kappa^2(u, cu)\, du.$$

Hence

$$Var(A_1) = O\left(\frac{1}{Sh^{2k}}\left(\frac{h^k}{n} + h^{k+2}\right)\right) = O\left(\frac{1}{Sh^k}\left(\frac{1}{n} + h^2\right)\right).$$

Since $h^2 >> \frac{1}{n}$, that $A_1 = o_P\left(\frac{1}{\sqrt{n}}\right)$ is implied by $Sh^{k-1}/n \to \infty$. □

D Proof of Theorem 3

Because of the linearity of the estimator, the proof is drastically simplified from Fan, Hu, and Truong (1994). Define $\kappa_s = \kappa\left(\frac{y^s}{h}\right)$ and $Z_s = \left(1, \frac{y^s}{h}\right)'$. Furthermore, let $m(y) = E(\eta | Y = y)$, $a_0 = m(0)$ and $b_0 = m'(0)$. Also let $\eta_s^* = y^s - a_0 - b_0 y^s$. Then one can write

$$\left(\hat{a} - a_0, h\left(\hat{b} - b_0\right)\right)' = \left(\sum_{s=1}^{S} Z_s Z_s \kappa_s\right)^{-1} \left(\sum_{s=1}^{S} Z_s \eta_s^* \kappa_s\right) = H^{-1} S.$$

Exact Identification Consider first H. Recall that

$$f_y(0) \to f_y^{\infty}(0) \equiv \frac{\pi(\theta_0)\det(W)^{1/2} e^{-\frac{1}{2}cZ'WZ}}{\rho_{\infty}(0)}.$$

Hence,

$$\frac{1}{Sh^k}H = (1 + o_P(1)) E\frac{1}{Sh^k}H,$$

where, for $C_{\kappa} = \int (1\ v)(1\ v)' \kappa(v)\, dv$,

$$E\frac{1}{Sh^k}H = \frac{1}{h^k}\int Z_s Z_s \kappa_s f(y^s)\, dy^s \to f_y^{\infty}(0) C_{\kappa}.$$

Now consider the bias and variance of S separately. Consider the bias first,

$$\begin{aligned} E\frac{1}{Sh^k h^2}S &= \frac{1}{h^{k+2}} EZ_s\kappa_s (E(\eta|y^s) - a_0 - b_0 y^s) = \frac{1}{h^{k+2}} EZ_s\kappa_s \left(\frac{1}{2}y_s' m''(0) y_s + O\left(y_s^3\right)\right) \\ &= \int (1\ v)' \frac{1}{2} v' m''(0) v\kappa_s(v)(1 + O(h)) f_y(vh)\, dv \\ &\longrightarrow f_y^{\infty}(0) \int (1\ v)' \frac{1}{2} v' m''(0) v\kappa_s(v)\, dv. \end{aligned}$$

Next consider the variance. Note that

$$\begin{aligned} Var\left(\frac{1}{\sqrt{Sh^k}}\sum_{s=1}^{S} Z_s\kappa_s\eta_s^*\right) =& \frac{1}{h^k} Var(Z_s\kappa_s\eta_s^*) \\ &= \frac{1}{h^k}\left[EVar(Z_s\kappa_s\eta_s^*|y_s) + VarE(Z_s\kappa_s\eta_s^*|y_s)\right]. \end{aligned}$$

For the first term,

$$\frac{m}{h^k} EVar(Z_s\kappa_s\eta_s^*|y_s) = \frac{m}{h^k} EZ_s Z_s'\kappa_s^2 Var(\eta^s|y^s) \to f_{\infty}(0)\mathcal{J}^{-1}\int (1\ v)(1\ v)'\kappa^2(v)\, dv.$$

For the second term,

$$\begin{aligned} &\frac{1}{h^k h^4} VarZ_s\kappa_s E(\eta_s^*|y_s) = \frac{1}{h^k h^4} VarZ_s\kappa_s \left(\frac{1}{2}y_s' m''(0) y_s + O\left(y_s^3\right)\right) \\ &\leq \frac{1}{h^k h^4} EZ_s Z_s'\kappa_s^2 \left(\frac{1}{2}y_s' m''(0) y_s + O\left(y_s^3\right)\right)^2 \\ &= \int (1\ v)(1\ v)'\kappa^2(v) \left(\left(\frac{1}{2}v' m''(0) v\right)^2 + O\left(h^2\right)\right) f_y(vh)\, dv \\ &\longrightarrow f_y^{\infty}(0) \int (1\ v)(1\ v)'\kappa^2(v) \left(\frac{1}{2}v' m''(0) v\right)^2 dv. \end{aligned}$$

These calculations show that $\frac{1}{Sh^k}H \to f_y^{\infty}(0) C_\kappa$ and that

$$\frac{1}{Sh^k}S = O_P\left(\frac{1}{\sqrt{Sh^k}}\left(\frac{1}{\sqrt{n}} + h^2\right) + h^2\right).$$

This implies the conclusion of the theorem since

$$\hat{\eta} - \bar{\eta} = O_P\left(\frac{1}{\sqrt{Sh^k}}\left(\frac{1}{\sqrt{n}} + h^2\right) + h^2\right).$$

Over Identification In this case the regressors Y_s are asymptotically collinear along a $d-k$ dimensional manifold with $1/\sqrt{n}$ variance. The coefficients in local linear regressions typically converge at a slower rate by the order of h than the intercept term. In this case, coefficients typically are slower by an order of $1/\sqrt{n}$, when $1/\sqrt{n} << h$. However, k linear combinations of the coefficients are only slower by an order of h.

Consider $m = n$ and consider the change of variable:

$$y_2 = g_2\left(g_1^{-1}(y_1)\right) + \frac{\Delta}{\sqrt{n}} = cy_1 + \left(g_2\left(g_1^{-1}(y_1)\right) - cy_1 + \frac{\Delta}{\sqrt{n}}\right).$$

Rewrite the local regression function as

$$\theta = a + b'y = a + w_1'd_1 + w_2'd_2 = a + w'd;$$

for $d_1 = (b_1 + c'b_2)$ and $d_2 = b_2/\sqrt{n}$ and $w = (w_1, w_2)$, $w_1 = y_1$,

$$w_2 = \sqrt{n}\left(g_2\left(g_1^{-1}(y_1)\right) - cy_1 + \frac{\Delta}{\sqrt{n}}\right).$$

Define $\kappa_s = \kappa\left(\frac{y^s}{h}\right)$ and $Z_s = \left(1, \frac{w_1^s}{h}, w_2^s\right)'$. Furthermore, let

$$m(y) = E(\eta | Y = y), \quad a_0 = m(0), \quad \text{and} \quad b_0 = m'(0).$$

Also let,

$$\eta_s^* = \theta^s - a_0 - b_0 y^s = \theta^s - a_0 - d_0 w^s.$$

Then, we can write

$$\left(\hat{a} - a_0, h\left(\hat{d}_1 - d_{10}\right), \hat{d}_2 - d_{20}\right)' = \left(\frac{1}{Sh^k}\sum_{s=1}^{S} Z_s Z_s \kappa_s\right)^{-1}\left(\frac{1}{Sh^k}\sum_{s=1}^{S} Z_s \eta_s^* \kappa_s\right) = H^{-1}S.$$

Consider first H. Note that

$$\kappa\left(\frac{y}{h}\right)=\kappa\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right).$$

Standard derivations imply $E\left(H-EH\right)^2=o(1)$ and where

$$\begin{aligned}EH=&\frac{1}{h^k}\int Z_sZ_s\kappa_sf\left(w^s\right)dw^s\\=&\frac{1}{h^k}\int\begin{pmatrix}1\\ \frac{w_1}{h}\\ w_2\end{pmatrix}\begin{pmatrix}1&\frac{w_1}{h}&w_2\end{pmatrix}\kappa\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)f\left(w_1,w_2\right)dw_1dw_2.\end{aligned}$$

By a change of variable $w_1=uh$, as $\sqrt{n}h\to\infty$,

$$EH\to\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\begin{pmatrix}1&u&w_2\end{pmatrix}\kappa\left(u,cu\right)f_w\left(0,w_2\right)dudw_2,$$

where w_2 is asymptotically normal and independent of w_1.

Next, consider bias and variance of S separately. Consider bias first. Note that $\eta\left(y\right)=\eta\left(y_1=w_1,y_2=cw_1+\frac{w_2}{\sqrt{n}}\right)$,

$$\begin{aligned}&E\frac{1}{h^2}S=\frac{1}{h^{k+2}}EZ_s\kappa_s\left(E\left(\eta|y^s\right)-a_0-b_0y^s\right)=\frac{1}{h^{k+2}}EZ_s\kappa_s\left(\frac{1}{2}y_s'm''\left(0\right)y_s+O\left(y_s^3\right)\right)\\=&\frac{1}{h^2}\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\kappa\left(u,cu+\frac{w_2}{\sqrt{n}h}\right)\times\\&\left[\frac{1}{2}\begin{pmatrix}uh\\ cuh+\frac{w_2}{\sqrt{n}}\end{pmatrix}'m''\left(0\right)\begin{pmatrix}uh\\ cuh+\frac{w_2}{\sqrt{n}}\end{pmatrix}+o\left(u^2h^2+\frac{w^2}{n}\right)\right]f_w\left(uh,w_2\right)dudw_2\\\to&\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\kappa\left(u,cu\right)\frac{1}{2}\begin{pmatrix}u\\ cu\end{pmatrix}'m''\left(0\right)\begin{pmatrix}u\\ cu\end{pmatrix}f_w\left(0,w_2\right)dudw_2.\end{aligned}$$

The variance also has two terms:

$$\begin{aligned}Var\left(\frac{1}{\sqrt{Sh^k}}\sum_{s=1}^{S}Z_s\kappa_s\eta_s^*\right)=&\frac{1}{h^k}Var\left(Z_s\kappa_s\eta_s^*\right)\\=&\frac{1}{h^k}\left[EVar\left(Z_s\kappa_s\eta_s^*|y_s\right)+VarE\left(Z_s\kappa_s\eta_s^*|y_s\right)\right].\end{aligned}$$

The first term in variance,

$$\frac{n}{h^k}EVar\left(Z_s\kappa_s\eta_s^*|y_s\right)=\frac{n}{h^k}EZ_sZ_s'\kappa_s^2Var\left(\eta^s|y^s\right)$$

$$=\frac{n}{h^k}\int\begin{pmatrix}1\\ \frac{w_1}{h}\\ w_2\end{pmatrix}\left(1\ \frac{w_1}{h}\ w_2\right)\kappa^2\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)$$

$$\times Var\left(\eta_s^*|y_1=w_1,y_2=cw_1+\frac{w_2}{\sqrt{n}}\right)f\left(w_1,w_2\right)dw_1dw_2$$

$$=\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\left(1\ u\ w_2\right)\kappa^2\left(u,cu+\frac{w_2}{\sqrt{n}}\right)\ n\ Var\left(\eta_s^*|y_1=uh,y_2=cuh+\frac{w_2}{\sqrt{n}}\right)f\left(uh,w_2\right)dudw_2$$

$$\to f_\infty\left(0\right)\mathcal{J}^{-1}\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\left(1\ u\ w_2\right)\kappa^2\left(u,cu\right)f\left(w_2\right)dudw_2.$$

The second term in variance,

$$\frac{1}{h^kh^4}VarZ_s\kappa_sE\left(\eta_s^*|y_s\right)=\frac{1}{h^kh^4}VarZ_s\kappa_s\left(\frac{1}{2}y_s'm''\left(0\right)y_s+O\left(y_s^3\right)\right)$$

$$\leq\frac{1}{h^kh^4}EZ_sZ_s'\kappa_s^2\left(\frac{1}{2}y_s'm''\left(0\right)y_s+O\left(y_s^3\right)\right)^2$$

$$=\frac{1}{h^kh^4}\int\begin{pmatrix}1\\ \frac{w_1}{h}\\ w_2\end{pmatrix}\left(1\ \frac{w_1}{h}\ w_2\right)\kappa^2\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)$$

$$\left(\frac{1}{2}\begin{pmatrix}w_1\\ cw_1+\frac{w_2}{\sqrt{n}}\end{pmatrix}'m''\left(0\right)\begin{pmatrix}w_1\\ cw_1+\frac{w_2}{\sqrt{n}}\end{pmatrix}+O\left(w_1^3+\frac{w_2^3}{n\sqrt{n}}\right)\right)^2f\left(w_1,w_2\right)dw_1dw_2$$

$$=\frac{1}{h^4}\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\left(1\ u\ w_2\right)\kappa^2\left(u,cu+\frac{w_2}{\sqrt{n}h}\right)$$

$$\left(\frac{1}{2}\begin{pmatrix}uh\\ cuh+\frac{w_2}{\sqrt{n}}\end{pmatrix}'m''\left(0\right)\begin{pmatrix}uh\\ cuh+\frac{w_2}{\sqrt{n}}\end{pmatrix}+o\left(u^2h^2+\frac{w_2^2}{n}\right)\right)^2f\left(uh,w_2\right)dudw_2.$$

As $h \to 0$, $n \to \infty$, this converges to

$$\int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} (1 \; u \; w_2) \kappa^2 (u, cu) \left(\frac{1}{2} \begin{pmatrix} u \\ cu \end{pmatrix}' m'' (0) \begin{pmatrix} u \\ cu \end{pmatrix} \right)^2 f (0, w_2) \, du dw_2.$$

Hence $H = O_P (1)$ and that

$$S = O_P \left(\frac{1}{\sqrt{Sh^k}} \left(\frac{1}{\sqrt{n}} + h^2 \right) + h^2 \right) = \hat{\eta} - \bar{\eta}.$$

Next, we briefly discuss the main issues in extending the discussion to the overidentified case with $m >> n$ but $m < \infty$. Note that now

$$y_1 = \hat{g}_1 (\theta) + \frac{\xi_1}{\sqrt{m}} \quad y_2 = \hat{g}_2 (\theta) + \frac{\xi_2}{\sqrt{m}}.$$

So that

$$\begin{aligned} y_2 = & \hat{g}_2 (\theta) + \frac{\xi_2}{\sqrt{m}} - g_2 \left(g_1^{-1} \left(\hat{g}_1 (\theta) - g_1 (\theta) + \frac{\xi_1}{\sqrt{m}} + g_1 (\theta) \right) \right) + g_2 \left(g_1^{-1} (y_1) \right) \\ = & c y_1 + g_2 \left(g_1^{-1} (y_1) \right) - c y_1 + \hat{g}_2 (\theta) - g_2 (\theta) + \frac{\xi_2}{\sqrt{m}} + O_P \left(\hat{g}_1 (\theta) - g_1 (\theta) + \frac{\xi_1}{\sqrt{m}} \right). \end{aligned}$$

The rate normalization for b_2 depends on the variation of $g_2 \left(g_1^{-1} (y_1) \right) - c y_1 = O (h^2)$, $\hat{g}_2 (\theta) - g_2 (\theta) = O_P \left(\frac{1}{\sqrt{n}} \right)$, $\hat{g}_1 (\theta) - g_1 (\theta) = O_P \left(\frac{1}{\sqrt{n}} \right)$, and $1/\sqrt{m}$, whichever prevails. The $O (h^2)$ and $O_P (1/\sqrt{m})$ terms only matter when the curvature of the moment conditions is such that $\hat{g}_j (\theta) - g_j (\theta)$ induces variations less than $O_P (1/\sqrt{n})$. □

E Proof of Theorem 4

We adapt and revise the local linear robust regression method of Fan, Hu, and Truong (1994) to our settings. Extensions to local polynomials are immediate. Let $\eta^s = a (\theta^s)$ for a known $a (\cdot) : R^k \to R$. The goal is to conduct inference on η_0. Define $\hat{\eta}_\tau = \hat{a}$, as in

$$\left(\hat{a}, \hat{b} \right) \equiv \arg \min_{a,b} \sum_{s=1}^{S} \rho_\tau (\eta^s - a - b y^s) \kappa \left(\frac{y^s}{h} \right).$$

For simplicity, confine to $m = n$. As preliminary preparations, define $\tilde{\eta} (y) = a \left(\tilde{\theta} (y) \right)$, where

$$\tilde{\theta} (y) = \arg \max_\theta \hat{Q} (\theta, y) = \frac{1}{2} (\hat{g} (\theta) - y)' \hat{W}_\theta (\hat{g} (\theta) - y).$$

Similarly, let $\eta(y) = a(\theta(y))$, where

$$\theta(y) = \arg\max_{\theta} Q(\theta, y) = \frac{1}{2}(g(\theta) - y)' \hat{W}_\theta (g(\theta) - y).$$

Recall that $f(\theta|Y = y) \propto \pi(\theta) e^{n\hat{Q}(\theta,y)}$, and that asymptotically as $n \to \infty$,

$$\theta|Y = y \xrightarrow{d} N\left(\tilde{\theta}(y), \mathcal{J}(y)^{-1}\right),$$

where $\mathcal{J}(y)^{-1} \equiv -\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}Q(\theta(y), y)$.

Alternatively,

$$\sqrt{n}\left(\theta - \tilde{\theta}(y)\right)|Y = y \overset{A}{\sim} N\left(0, \mathcal{J}(y)^{-1}\right).$$

Hence, $\sqrt{n}(\eta - \tilde{\eta}(y))|Y = y \overset{A}{\sim} N\left(0, \partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y\right)$, or $\eta|Y = y \overset{A}{\sim} N\left(\tilde{\eta}(y), \frac{1}{n}\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y\right)$. Define then,

$$m(y) \equiv Q_\tau(\eta|Y = y) \sim Q_\tau\left(N\left(\tilde{\eta}(y), \frac{1}{n}\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y\right)\right) = \tilde{\eta}(y) + \frac{1}{\sqrt{n}}\sqrt{\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y}Z_\tau.$$

This implies that

$$m'(y) \sim \frac{\partial}{\partial y}\tilde{\eta}(y) + \frac{1}{\sqrt{n}}\frac{\partial}{\partial y}\sqrt{\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y}Z_\tau$$

and

$$m''(y) \sim \frac{\partial^2}{\partial y\partial y'}\tilde{\eta}(y) + \frac{1}{\sqrt{n}}\frac{\partial^2}{\partial y\partial y'}\sqrt{\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y}Z_\tau.$$

Exact Identification: $d = k$. Let $a_0 = m(0)$, $b_0 = m'(0)$, and $Z_s = \left(1, \frac{y_s}{h}\right)$. Define $\theta = \sqrt{n}\sqrt{Sh^k}\left(\hat{a} - a_0, h\left(\hat{b} - b_0\right)\right)$. Let $\eta_s^* = \eta_s - a_0 - b_0'y_s$, and $\kappa_s = \kappa\left(\frac{y_s}{h}\right)$. Then $\hat{\theta}$ minimizes

$$G_S(\theta) = \sqrt{n}\sum_{s=1}^{S}\left(\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau(\eta_s^*)\right)\kappa_s.$$

Consider now the decomposition that

$$G_S(\theta) = E(G_S(\theta)|Y) + \left(Sh^k\right)^{-1/2}\sum_{s=1}^{S}\left(\rho_\tau'(\eta_s^*) Z_s'\kappa_s - E(\rho_\tau'(\eta_s^*)|y_s) Z_s'\kappa_s\right)\theta + R_S(\theta),$$

where we have defined $\rho_\tau'(\cdot) = \tau - 1(\cdot \leq 0)$.

Assume that the kernel function $\kappa(\cdot)$ has bounded support (by M for example). First, note that

$$m(y_s) = a_0 + b_0 y_s + \frac{1}{2} y_s' m''(0) y_s + o_P\left(h^2\right).$$

Now write

$$\begin{aligned} E(G_S(\theta)|Y) =& \sqrt{n} \sum_{s=1}^{S} E\left[\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau(\eta_s^*)|y_s\right]\kappa_s \\ =& \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} E\left(\rho_\tau'(\eta_s^*)|y_s\right) Z_s' \kappa_s \theta \\ &+ \frac{1}{2}\frac{1}{\sqrt{n}Sh^k}\theta' \sum_{s=1}^{S} E\left(\rho_\tau''(\eta_s^*)|y_s\right) Z_s Z_s' \kappa_s \theta \left(1 + o_P(1)\right). \end{aligned}$$

Consider first

$$\begin{aligned} E\left(\rho_\tau''(\eta_s^*)|y_s\right) =& f_\eta\left(a_0 + b_0 y_s | y_s\right) \approx \phi\left(a_0 + b_0 y_s \middle| \tilde{\eta}(y), \frac{1}{n}\partial\eta_y' \mathcal{J}_y^{-1} \partial\eta_y\right) \\ =& \frac{\sqrt{n}}{\sqrt{\partial\eta_y' \mathcal{J}_y^{-1}\partial\eta_y}} \phi\left(\frac{a_0 + b_0 y_s - \tilde{\eta}(y)}{\frac{1}{\sqrt{n}}\sqrt{\partial\eta_y' \mathcal{J}_y^{-1}\partial\eta_y}}\right) \\ =& \frac{\sqrt{n}}{\sqrt{\partial\eta_y' \mathcal{J}_y^{-1}\partial\eta_y}} \phi\left(\frac{a_0 + b_0 y_s - m(y_s)}{\frac{1}{\sqrt{n}}\sqrt{\partial\eta_y' \mathcal{J}_y^{-1}\partial\eta_y}} + Z_\tau\right). \end{aligned}$$

As $y_s \to 0$, $\sqrt{n}\left(a_0 + b_0 y_s - m(y_s)\right) \to 0$, and

$$\frac{1}{\sqrt{n}} E\left(\rho_\tau''(\eta_s^*)|y_s\right) \to \frac{1}{\sqrt{\partial\eta' \mathcal{J}^{-1}\partial\eta}} \phi(Z_\tau),$$

as $h \to 0$ and $n \to \infty$. We also have $y_s^2 << \sqrt{n}$ with bounded kernels, since $h << n^{-1/4}$. Also recall that,

$$f_y(0) \to f_y^\infty(0) \equiv \frac{\pi(\theta_0) \det(W)^{1/2} e^{-\frac{1}{2}cZ'WZ}}{\rho_\infty(0)}.$$

Therefore

$$\frac{1}{\sqrt{n}Sh^k} \sum_{s=1}^{S} E\left(\rho_\tau''(\eta_s^*)|y_s\right) Z_s Z_s' \kappa_s \to \frac{1}{\sqrt{n}} \int \frac{1}{h^k} E\left(\rho_\tau''(\eta_s^*)|y_s\right) Z_s Z_s' \kappa_s f(y_s)\, dy_s$$

By a change of variable $y_s = vh$, this converges to, for $C_\kappa = \int (1\ v)(1\ v)' \kappa(v)\, dv$,

$$H = C_\kappa \frac{1}{\sqrt{\partial\eta' \mathcal{J}^{-1}\partial\eta}} \phi(Z_\tau) f_y^\infty(0).$$

Hence we can write, for θ in a compact set,

$$G_S(\theta) = \frac{1}{2}\theta' H\theta + \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} \rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s\theta + R_n(\theta) + o_P(1).$$

Next we show that $R_S(\theta) = o_P(1)$. Since $ER_S(\theta) = 0$, it suffices to bound $ER_S(\theta)^2$,

$$\begin{aligned}
ER_S(\theta)^2 \leq & nSE\left[\left(\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau\left(\eta_s^*\right)\right) - \frac{1}{\sqrt{n}\sqrt{Sh^k}}\rho'_\tau\left(\eta_s^*\right)\theta' Z_s\right]^2 \kappa_s^2 \\
\leq & nSE\left[1\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s\right)\left(\frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s\right)^2\right]\kappa_s^2 \\
\leq & E\frac{1}{h^k}\left(\theta' Z_s\right)^2 \kappa_s^2 P\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s | y_s\right) = O\left(\frac{1}{\sqrt{Sh^k}}\right) = o(1)
\end{aligned}$$

as long as $Sh^k \to \infty$. The above relation follows from $f_y(0) = O(1)$, that $\theta' Z_s \leq M$ for M being the support of κ_s, and $P\left(|\eta_s^*| \leq \frac{x}{\sqrt{n}}|y_s\right) = O(x)$. The last relation in turn follows from $\sqrt{n}\eta_s^* \overset{A}{\sim} N\left(0, \partial\eta_y' \mathcal{J}_y^{-1}\partial\eta_y\right)$. Next, if we can show that

$$W_s \equiv \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} \rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s$$

is stochastically bounded and asymptotically normal, then $\hat{\theta} - H^{-1}W_S = o_P(1)$. We check both $Var(W_S)$ and $E(W_S)$.

$$Var(W_S) = \frac{1}{h^k} Var\left(\rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s\right) = \frac{1}{h^k}\left[E_{y_s} Var\left(\rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s | y_s\right) + Var_{y_s} E\left(\rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s | y_s\right)\right].$$

Recall that $\rho'_\tau\left(\eta_s^*\right) = \tau - 1\left(\eta_s - a_0 - b_0 y_s \leq 0\right)$, it can be calculated that

$$\begin{aligned}
E_{y_s} Var\left(\rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s | y_s\right) = & E_{y_s} Z_s Z_s'\kappa_s^2 Var\left(\tau - 1\left(\eta_s \leq a_0 + b_0 y_s\right) | y_s\right) \\
= & \int Z_s Z_s \kappa_s^2 P\left(\eta_s \leq a_0 - b_0 y_s | y_s\right)\left(1 - P\left(\eta_s \leq a_0 - b_0 y_s | y_s\right)\right) f\left(y_s\right) dy_s.
\end{aligned}$$

Hence by the usual change of variable $y_s = vh$, for $\bar{C}_\kappa = \int (1\ v)(1\ v)' \kappa^2(v)\, dv$,

$$\frac{1}{h^k} E_{y_s} Var\left(\rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s | y_s\right) \to \tau(1-\tau)\bar{C}_\kappa f_y^\infty(0).$$

Next,

$$\begin{aligned}
Var_{y_s} E\left(\rho'_\tau\left(\eta_s^*\right) Z_s'\kappa_s | y_s\right) = & Var_{y_s} Z_s\kappa_s E\left(\tau - 1\left(\eta_s \leq a_0 + b_0 y_s\right) | y_s\right) \\
\approx & Var_{y_s}\left[Z_s\kappa_s f_\eta\left(Q_\tau\left(\eta | y_s\right)\right)\frac{1}{2}m''(0)\, y_s^2\right].
\end{aligned}$$

Hence,

$$\frac{1}{h^k n h^4} Var_{y_s} E\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) \to 1/4 m''\left(0\right)^2 \tilde{C}_\kappa \frac{1}{\partial\eta' \mathcal{J}^{-1} \partial\eta} \phi^2\left(Z_\tau\right) f_y^\infty\left(0\right),$$

where $\tilde{C}_\kappa = \int \left(1\ v\right)\left(1\ v\right)' v^4 \kappa^2\left(v\right) dv$. In order that the second part of the variance does not dominate, we require that $nh^4 \to 0$.

Consider finally the bias term:

$$\begin{aligned} E\frac{1}{\sqrt{n}\sqrt{Sh^k}h^2} W_s =& E\frac{1}{\sqrt{n}h^k h^2}\left(\tau - 1\left(\eta_s \leq a_0 + b_0 y_s\right)|y_s\right) Z_s \kappa_s \\ =& \frac{1}{h^k h^2} E Z_s \kappa_s \frac{1}{\sqrt{n}} f_\eta\left(Q_\tau\left(\eta | y_s\right)\right) \frac{1}{2} m''\left(0\right) y_s^2 \left(1 + o_P\left(1\right)\right) \\ \to& \frac{1}{2} m''\left(0\right) \frac{1}{\sqrt{\partial\eta' \mathcal{J}^{-1} \partial\eta}} f_y^\infty\left(0\right) \phi\left(Z_\tau\right) \int \left(1\ v\right)' \kappa\left(v\right) v^2 dv. \end{aligned}$$

Hence we conclude when $nh^4 \to 0$, $W_S = O_P\left(1 + \sqrt{nSh^k}h^2\right)$, so that $\hat{\theta} = -H^{-1} W_S + o_P\left(1\right) = O_P\left(1 + \sqrt{nSh^k}h^2\right)$. This implies that $\hat{a} - a_0 = O_P\left(\frac{1}{\sqrt{nSh^k}} + h^2\right)$. Whenever $nSh^{2k+4} \to 0$ so that the bias is sufficiently small, $\hat{\theta} \xrightarrow{d} N\left(0, \tau\left(1-\tau\right) H^{-1} \bar{C}_\kappa f_y^\infty\left(0\right) H^{-1}\right)$. However, the asymptotic normality of $\hat{\theta}$ is of less importance than the rate condition. When $\sqrt{n}h^2 \to 0$ and $Sh^k \to \infty$ (which neccesitates that $S >> n^{k/4}$), $\hat{\theta} = o_P\left(1/\sqrt{n}\right)$. In this case the simulated posterior quantiles provide a valid inference method, as long as the infeasible posterior quantiles $\bar{\eta}_\tau$ are valid. For example,

$$P\left(\sqrt{n}\left(\hat{\eta}_\tau - \eta_0\right) \leq 0\right) = P\left(\sqrt{n}\left(\bar{\eta}_\tau - \eta_0\right) + o_P\left(1\right) \leq 0\right) = P\left(\sqrt{n}\left(\hat{\eta}_\tau - \eta_0\right) \leq 0\right) + o_P\left(1\right).$$

The same proof can be adapted for the local linear estimator of the posterior mean. In that case instead of $l\left(x\right) = \rho_\tau\left(x\right)$, let $l\left(x\right) = \left(x\right)^2$. Then $l'\left(x\right) = 2x$, and $l''\left(x\right) = 2$. A different normalization of the objective function should be used however. Define now

$$G_S\left(\theta\right) = n \sum_{s=1}^S \left(l\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{nSh^k}}\right) - l\left(\eta_s^*\right)\right) \kappa_s.$$

Then a similar sequence of arguments will go through, now with $W_S \equiv \sqrt{n}\frac{1}{\sqrt{Sh^k}} \sum_{s=1}^S l'\left(\eta_s^*\right) Z_s \kappa_s$.

For the quantile case, furthermore, the above arguments for proving $\hat{a} - a_0 = O_P\left(\frac{1}{\sqrt{nSh^k}} + h^2\right)$ strictly speaking only work for the case where $nSh^{2k+4} = O\left(1\right)$. In the other case when $nSh^{2k+4} \to \infty$ so that the bias term dominates, a similar sequence of arguments can be

followed, but the objective function needs to be normalized differently. In this case, let $\theta = h^{-2}\left(\left(a-a_0\right), h\left(b-b_0\right)\right)$, and redefine

$$\hat{G}_S\left(\theta\right) = \frac{1}{h^4\sqrt{n}Sh^k}\sum_{s=1}^{S}\left(\rho_\tau\left(\eta_s^* - h^2\theta\right) - \rho_\tau\left(\eta_s^*\right)\right)\kappa_s.$$

Then one can show that in

$$G_S\left(\theta\right) = E\left(G_S\left(\theta\right)|Y\right) + \frac{1}{h^2\sqrt{n}Sh^k}\sum_{s=1}^{S}\left(\rho_\tau'\left(\eta_s^*\right)Z_s'\kappa_s - E\left(\rho_\tau'\left(\eta_s^*\right)|y_s\right)Z_s'\kappa_s\right)\theta + R_S\left(\theta\right),$$

it holds that

$$E\left(G_S\left(\theta\right)|Y\right) = \frac{1}{2}\theta' H\theta\left(1+o_P\left(1\right)\right) + \frac{1}{h^2\sqrt{n}Sh^k}\sum_{s=1}^{S}E\left(\rho_\tau'\left(\eta_s^*\right)|y_s\right)Z_s'\kappa_s\theta.$$

Hence

$$G_S\left(\theta\right) = \frac{1}{2}\theta' H\theta + \frac{1}{h^2\sqrt{n}Sh^k}\sum_{s=1}^{S}\rho_\tau'\left(\eta_s^*\right)Z_s'\kappa_s\theta + R_n\left(\theta\right) + o_P\left(1\right).$$

Then in combination of the result that

$$\frac{1}{h^2\sqrt{n}Sh^k}\sum_{s=1}^{S}\rho_\tau'\left(\eta_s^*\right)Z_s'\kappa_s \xrightarrow{p} C$$

for some finite C, we will have that $\hat{a}-a = O_P\left(h^2\right)$ when $h^2 >> \sqrt{nSh^k}$.

Overidentification Let $a_0 = m\left(0\right)$, $b_0 = m'\left(0\right)$, define a similar reparameterization analogous to the mean case:

$$d_1 = \left(b_1 + c'b_2\right), \quad d_2 = b_2/\sqrt{n}, \quad w_1 = y_1, \quad w_2 = \sqrt{n}\left(g_2\left(g_1^{-1}\left(y_1\right)\right) - cy_1 + \frac{\Delta}{\sqrt{n}}\right).$$

Also let $\kappa_s = \kappa\left(\frac{y_s}{h}\right)$ and $Z_s = \left(1, \frac{w_1^s}{h}, w_2^s\right)$. Define $\theta = \sqrt{n}\sqrt{Sh^k}\left(\hat{a}-a_0, h\left(\hat{d}_1 - d_{10}\right), \left(\hat{d}_2 - d_{20}\right)\right)$,

$$\eta_s^* = \eta_s - a_0 - b_0'y_s = \eta_s - a_0 - d_0'w_s.$$

Then $\hat{\theta}$ minimizes Koenker and Bassett (1978)'s check function

$$G_S\left(\theta\right) = \sqrt{n}\sum_{s=1}^{S}\left(\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau\left(\eta_s^*\right)\right)\kappa_s.$$

Consider the decomposition, for $\rho'_\tau(\cdot) = \tau - 1(\cdot \leq 0)$,

$$G_S(\theta) = E(G_S(\theta)|Y) + \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} \left(\rho'_\tau(\eta_s^*) Z'_s \kappa_s - E\left(\rho'_\tau(\eta_s^*)|y_s\right) Z'_s \kappa_s\right) \theta + R_S(\theta).$$

Noting that

$$m(y_s) = a_0 + b_0 y_s + \frac{1}{2} y'_s m''(0) y_s + o_P\left(h^2 + \frac{1}{\sqrt{n}}\right).$$

We write

$$E(G_S(\theta)|Y) = \sqrt{n} \sum_{s=1}^{S} E\left[\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau(\eta_s^*)|y_s\right] \kappa_s$$

$$= \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} E\left(\rho'_\tau(\eta_s^*)|y_s\right) Z'_s \kappa_s \theta + \frac{1}{2} \frac{1}{\sqrt{n} Sh^k} \theta' \sum_{s=1}^{S} E\left(\rho''_\tau(\eta_s^*)|y_s\right) Z_s Z'_s \kappa_s \theta (1 + o_P(1)).$$

Consider first

$$\begin{aligned} E\left(\rho''_\tau(\eta_s^*)|y_s\right) =& f_\eta(a_0 + b_0 y_s|y_s) \approx \phi\left(a_0 + b_0 y_s \middle| \tilde{\eta}(y), \frac{1}{n} \partial\eta'_y \mathcal{J}_y^{-1} \partial\eta_y\right) \\ =& \frac{\sqrt{n}}{\sqrt{\partial\eta'_y \mathcal{J}_y^{-1} \partial\eta_y}} \phi\left(\frac{a_0 + b_0 y_s - \tilde{\eta}(y)}{\frac{1}{\sqrt{n}}\sqrt{\partial\eta'_y \mathcal{J}_y^{-1} \partial\eta_y}}\right) \\ =& \frac{\sqrt{n}}{\sqrt{\partial\eta'_y \mathcal{J}_y^{-1} \partial\eta_y}} \phi\left(\frac{a_0 + b_0 y_s - m(y_s)}{\frac{1}{\sqrt{n}}\sqrt{\partial\eta'_y \mathcal{J}_y^{-1} \partial\eta_y}} + Z_\tau\right). \end{aligned}$$

As $y_s \to 0$, $a_0 + b_0 y_s - m(y_s) = O_P(h^2) = o_P(1/\sqrt{n})$, and

$$\frac{1}{\sqrt{n}} E\left(\rho''_\tau(\eta_s^*)|y_s\right) \to \frac{1}{\sqrt{\partial\eta' \mathcal{J}^{-1} \partial\eta}} \phi(Z_\tau).$$

Change variable $y = (y_1,\ y_2) = \left(w_1,\ cw_1 + \frac{w_2}{\sqrt{n}}\right)$

$$\frac{1}{\sqrt{n} Sh^k} \sum_{s=1}^{S} E\left(\rho''_\tau(\eta_s^*)|y_s\right) Z_s Z'_s \kappa_s \to \frac{1}{\sqrt{n}} \int\int \frac{1}{h^k} E\left(\rho''_\tau(\eta_s^*)|y_s\right) Z_s Z'_s \kappa_s f(y_s)\, dy_s$$

$$= (1 + o(1))\ \frac{1}{\sqrt{\partial\eta' \mathcal{J}^{-1} \partial\eta}} \phi(Z_\tau) \frac{1}{h^k} \int\int \begin{pmatrix} 1 \\ \frac{w_1}{h} \\ w_2 \end{pmatrix} \left(1\ \ \frac{w_1}{h}\ \ w_2\right) \kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) dF(w_1, w_2)$$

$$\to H = \frac{1}{\sqrt{\partial\eta' \mathcal{J}^{-1} \partial\eta}} \phi(Z_\tau) \int\int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} (1\ u\ w_2)\, \kappa(u, cu)\, f(0, w_2)\, du dw_2.$$

Then for θ in a compact set,

$$G_S(\theta) = \frac{1}{2}\theta' H\theta + \left(\tau_n S h^k\right)^{-1/2} \sum_{s=1}^{S} \rho'_\tau\left(\eta_s^*\right) Z_s' \kappa_s \theta + R_S(\theta) + o_P(1).$$

Next $R_S(\theta) = o_P(1)$. Since $ER_S(\theta) = 0$, and

$$\begin{aligned} ER_S(\theta)^2 \leq & nSE\left[\left(\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau\left(\eta_s^*\right)\right) - \frac{1}{\sqrt{n}\sqrt{Sh^k}}\rho'_\tau\left(\eta_s^*\right)\theta' Z_s\right]^2 \kappa_s^2 \\ \leq & nSE\left[1\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s\right)\left(\frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s\right)^2\right]\kappa_s^2 \\ \leq & E\frac{1}{h^k}\left(\theta' Z_s\right)^2 \kappa_s^2 P\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s | y_s\right) = O\left(\frac{1}{\sqrt{Sh^k}}\right) = o(1). \end{aligned}$$

In the above $P\left(|\eta_s^*| \leq \frac{x}{\sqrt{n}}|y_s\right) = O(x)$, $\theta' Z_s = O_P(1)$. This is verified by change of variable in integration.

If we can show that

$$W_s \equiv \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} \rho'_\tau\left(\eta_s^*\right) Z_s' \kappa_s = O_P(1),$$

then $\hat{\theta} - H^{-1}W_S = o_P(1)$. For this purpose we check both $Var(W_S)$ and $E(W_S)$,

$$\begin{aligned} Var(W_S) = & \frac{1}{h^k} Var\left(\rho'_\tau\left(\eta_s^*\right) Z_s' \kappa_s\right) \\ = & \frac{1}{h^k}\left[E_{y_s} Var\left(\rho'_\tau\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) + Var_{y_s} E\left(\rho'_\tau\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right)\right]. \end{aligned}$$

Recall that $\rho'_\tau\left(\eta_s^*\right) = \tau - 1\left(\eta_s - a_0 - b_0 y_s \leq 0\right)$,

$$\begin{aligned} & \frac{1}{h^k} E_{y_s} Var\left(\rho'_\tau\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) = \frac{1}{h^k} E_{y_s} Z_s Z_s' \kappa_s^2 Var\left(\tau - 1\left(\eta_s - a_0 - b_0 y_s\right) | y_s\right) \\ = & \frac{1}{h^k}\int Z_s Z_s \kappa_s^2 P\left(\eta_s \leq a_0 + b_0 y_s | y_s\right)\left(1 - P\left(\eta_s \leq a_0 + b_0 y_s | y_s\right)\right) f\left(y_s\right) dy_s \\ = & (1 + o(1))\, \tau(1-\tau)\frac{1}{h^k}\int \begin{pmatrix} 1 \\ \frac{w_1}{h} \\ w_2 \end{pmatrix}\left(1 \;\; \frac{w_1}{h} \;\; w_2\right)\kappa^2\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) f_w(w_1, w_2)\, du dw_2 \\ \to & \tau(1-\tau)\int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix}\kappa^2(u, cu)\, f_w(0, w_2)\, du dw_2. \end{aligned}$$

The second term in variance, under the condition that $nh^4 \to 0$,

$$\begin{aligned} Var_{y_s} E\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) =& Var_{y_s} Z_s \kappa_s E\left(\tau - 1\left(\eta_s \leq a_0 + b_0 y_s\right) | y_s\right) \\ \approx& Var_{y_s}\left[Z_s \kappa_s f_\eta\left(Q_\tau\left(\eta | y_s\right)\right) \frac{1}{2} y_s' m''(0)\, y_s\right]. \end{aligned}$$

This converges to

$$\begin{aligned} & \frac{1}{h^k n h^4} Var_{y_s} E\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) \\ (1+o(1)) & \frac{1}{\partial \eta' \mathcal{J}^{-1} \partial \eta} \phi^2\left(Z_\tau\right) \frac{1}{h^{k+4}} \int\int \begin{pmatrix} 1 \\ \frac{w_1}{h} \\ w_2 \end{pmatrix} \begin{pmatrix} 1 & \frac{w_1}{h} & w_2 \end{pmatrix} \kappa^2\left(\frac{w_1}{h}, \frac{c w_1}{h} + \frac{w_2}{\sqrt{n} h}\right) \\ \times & \left[\begin{pmatrix} w_1 \\ c w_1 + \frac{w_2}{\sqrt{n}} \end{pmatrix}' \frac{1}{2} m''(0) \begin{pmatrix} w_1 \\ c w_1 + \frac{w_2}{\sqrt{n}} \end{pmatrix}\right]^2 dF\left(w_1, w_2\right) \\ \to & \frac{1}{\partial \eta' \mathcal{J}^{-1} \partial \eta} \phi^2\left(Z_\tau\right) \int\int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \begin{pmatrix} 1 & u & w_2 \end{pmatrix} \kappa^2(u, cu) \\ \times & \left[\begin{pmatrix} u \\ cu \end{pmatrix}' \frac{1}{2} m''(0) \begin{pmatrix} u \\ cu \end{pmatrix}\right]^2 f_w(0, w_2)\, dw_1 dw_2. \end{aligned}$$

Finally, consider the bias term

$$
\begin{aligned}
E\frac{1}{\sqrt{n}\sqrt{Sh^k}h^2}W_s &= E\frac{1}{\sqrt{n}h^k h^2}\left(\tau - P\left(\eta_s \leq a_0 + b_0 y_s\right)|y_s\right) Z_s\kappa_s \\
&= \frac{1}{h^k h^2} E Z_s\kappa_s \frac{1}{\sqrt{n}} f_\eta\left(Q_\tau\left(\eta|y_s\right)\right) y_s' \frac{1}{2} m''\left(0\right) y_s \left(1 + o_P\left(1\right)\right) \\
&= (1+o(1))\frac{\phi\left(Z_\tau\right)}{\sqrt{\partial\eta'\mathcal{J}^{-1}\partial\eta}}\frac{1}{h^k h^2}\int\int \begin{pmatrix} 1 \\ \frac{w_1}{h} \\ w_2 \end{pmatrix} \kappa^2\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) \\
&\quad \begin{pmatrix} w_1 \\ cw_1 + \frac{w_2}{\sqrt{n}} \end{pmatrix}' \frac{1}{2} m''\left(0\right) \begin{pmatrix} w_1 \\ cw_1 + \frac{w_2}{\sqrt{n}} \end{pmatrix} dF(w_1, w_2) \\
\times \quad &\to \frac{\phi\left(Z_\tau\right)}{\sqrt{\partial\eta'\mathcal{J}^{-1}\partial\eta}}\int\int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \kappa^2\left(u, cu\right) \\
\times \quad &\quad \begin{pmatrix} u \\ cu \end{pmatrix}' \frac{1}{2} m''\left(0\right) \begin{pmatrix} u \\ cu \end{pmatrix}' f_w(0, w_2) du dw_2.
\end{aligned}
$$

When $nh^4 \to 0$, $W_S = O_P\left(1 + \sqrt{nSh^k}h^2\right)$, so that

$$\hat{\theta} = -H^{-1}W_S + o_P\left(1\right) = O_P\left(1 + \sqrt{nSh^k}h^2\right).$$

This implies that $\hat{a} - a_0 = O_P\left(\frac{1}{\sqrt{nSh^k}} + h^2\right)$. To conclude, when $\sqrt{n}h^2 \to 0$ and $Sh^k \to \infty$ (which neccesitates that $S >> n^{k/4}$), $\hat{\theta} = o_P\left(1/\sqrt{n}\right)$, the simulated posterior quantiles provide a valid inference method.

Similar to the exact identification case, strictly speaking, the previous normalization only works when $nSh^{2k+4} = O\left(1\right)$. In the case when $nSh^{2k+4} \to \infty$ so that the bias term dominates, the objective function also needs to be normalized differently, by letting $\theta = h^{-2}\left(\left(a - a_0\right), h\left(d_1 - d_{10}\right), \left(d_2 - d_{20}\right)\right)$ and redefining

$$G_S\left(\theta\right) = \frac{1}{h^4\sqrt{n}Sh^k}\sum_{s=1}^{S}\left(\rho_\tau\left(\eta_s^* - h^2\theta' Z_s\right) - \rho_\tau\left(\eta_s^*\right)\right)\kappa_s.$$

F Proof of Theorem 5

We focus on the exact identification case. The arguments for the overidentified case is similar to local linear regressions with properly defined notations. Recall that $\kappa_s = \kappa\left(\frac{y^s}{h}\right)$ and $m(y) = E(\eta|Y=y)$. Define $b_u = h^{[u]}(\beta_u - \beta_u^0)$, and $b = (b_u, u \in A)$. Also let $Z_s^u = y_s^u h^{[-u]}$, and that $Z_s^A = (Z_s^u, u \in A)$. Also, let $\eta_s^* = \eta_s - \beta_0' y_s^A$.

Mean regression We can now write

$$\hat{b} = \left(\sum_{s=1}^S Z_s^A Z_s^A \kappa_s\right)^{-1}\left(\sum_{s=1}^S Z_s^A \eta_s^* \kappa_s\right) = H^{-1}S.$$

Consider H first, recall that $f_y(0) \to f_y^\infty(0)$ and $\frac{1}{Sh^k}H = (1+o_P(1))E\frac{1}{Sh^k}H$. Then for

$$C_\kappa = \int v_A v_A' \kappa(v)\, dv, \qquad v_A = (v^u = v_1^{u_1}\ldots v_d^{u_d}, u \in A),$$
$$E\frac{1}{Sh^k}H = \frac{1}{h^k}\int Z_s^A Z_s^A \kappa_s f(y^s)\, dy^s \to f_y^\infty(0) C_\kappa.$$

Now consider the bias and variance of S separately. Consider the bias first. Note that

$$m(y_s) - \sum_{u\in A}\beta_u y_s^u = \sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u} m(0) + o_P\left(\frac{1}{\sqrt{n}}\right) + O\left(h^{p+2}\right).$$

Then

$$\begin{aligned}
&E\frac{1}{Sh^k h^{p+1}}S \\
&\quad = \frac{1}{h^{k+p+1}} E\left(\frac{y_s^u}{h^{[u]}}, u\in A\right)\kappa_s\left(\frac{y_s}{h}\right)\left[o_P\left(\frac{1}{\sqrt{n}}\right) + O\left(h^{p+2}\right) + \sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u} m(0)\right] \\
&\quad = \int v_A \kappa(v)\left[o_P\left(\frac{1}{\sqrt{n}h^{p+1}}\right) + O_P(h) + \sum_{[u]=p+1} v^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u} m(0)\right] f_y(vh)\, dv \\
&\quad \to f_y^\infty(0)\int v_A \kappa(v) \sum_{[u]=p+1} v^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u} m(0)\, dv.
\end{aligned}$$

Next consider the variance. Note that

$$\begin{aligned}
Var\left(\frac{1}{\sqrt{Sh^k}}\sum_{s=1}^S Z_s^A \kappa_s \eta_s^*\right) &= \frac{1}{h^k} Var\left(Z_s^A \kappa_s \eta_s^*\right) \\
&= \frac{1}{h^k}\left[EVar\left(Z_s^A \kappa_s \eta_s^* | y_s\right) + VarE\left(Z_s^A \kappa_s \eta_s^* | y_s\right)\right].
\end{aligned}$$

For the first term,

$$\frac{n}{h^k}EVar\left(Z_s^A\kappa_s\eta_s^*|y_s\right)=\frac{n}{h^k}EZ_s^AZ_s^{A'}\kappa_s^2Var\left(\eta^s|y^s\right)\to f_\infty\left(0\right)\mathcal{J}^{-1}\int v_Av_A'\kappa^2\left(v\right)dv.$$

For the second term,

$$\begin{aligned}
&\frac{1}{h^kh^{2(p+1)}}VarZ_s^A\kappa_sE\left(\eta_s^*|y_s\right)\\
&=\frac{1}{h^kh^{2(p+1)}}VarZ_s^A\kappa_s\left[o_P\left(\frac{1}{\sqrt{n}}\right)+O\left(h^{p+2}\right)+\sum_{[u]=p+1}y_s^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}m\left(0\right)\right]\\
&\leq\frac{1}{h^kh^{2(p+1)}}EZ_s^AZ_s^{A'}\kappa_s^2\left[o_P\left(\frac{1}{\sqrt{n}}\right)+O\left(h^{p+2}\right)+\sum_{[u]=p+1}y_s^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}m\left(0\right)\right]^2\\
&=\int v_Av_A'\kappa^2\left(v\right)\left(\left(\sum_{[u]=p+1}v^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}m\left(0\right)\right)^2+O_P\left(\frac{1}{nh^{2(p+1)}}+h^2\right)\right)f_y\left(vh\right)dv\\
&\longrightarrow f_y^\infty\left(0\right)\int v_Av_A'\kappa^2\left(v\right)\left(\sum_{[u]=p+1}v^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}m\left(0\right)\right)^2dv.
\end{aligned}$$

This implies the conclusion of the theorem since

$$\hat{\eta}-\bar{\eta}=O_P\left(\frac{1}{Sh^k}S\right)=O_P\left(\frac{1}{\sqrt{Sh^k}}\left(\frac{1}{\sqrt{n}}+h^{p+1}\right)+h^{p+1}\right).$$

Quantile Regression Define

$$\theta=\sqrt{n}\sqrt{Sh^k}b=\sqrt{n}\sqrt{Sh^k}\left(b_u,u\in A\right)=\sqrt{n}\sqrt{Sh^k}\left(h^{[u]}\left(\beta_u-\beta_u^0\right),u\in A\right).$$

Note $\eta_s^*=\eta_s-\beta_0'y_s^A=\eta_s-\theta_0'Z_s^A$. Then $\hat{\theta}$ minimizes

$$G_S\left(\theta\right)=\sqrt{n}\sum_{s=1}^S\left(\rho_\tau\left(\eta_s^*-\frac{\theta'Z_s^A}{\sqrt{n}\sqrt{Sh^k}}\right)-\rho_\tau\left(\eta_s^*\right)\right)\kappa_s.$$

Consider now the decomposition that

$$G_S\left(\theta\right)=E\left(G_S\left(\theta\right)|Y\right)+\left(Sh^k\right)^{-1/2}\sum_{s=1}^S\left(\rho_\tau'\left(\eta_s^*\right)Z_s^{A'}\kappa_s-E\left(\rho_\tau'\left(\eta_s^*\right)|y_s\right)Z_s^{A'}\kappa_s\right)\theta+R_S\left(\theta\right),$$

where we have defined $\rho_\tau'\left(\cdot\right)=\tau-1\left(\cdot\leq0\right)$. First, note that

$$m\left(y_s\right)=\theta'Z_s^A+\sum_{[u]=p+1}y_s^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}m\left(0\right)+o_P\left(\frac{1}{\sqrt{n}}\right)+O\left(h^{p+2}\right).$$

Now write

$$E\left(G_S\left(\theta\right)|Y\right)=\sqrt{n}\sum_{s=1}^{S}E\left[\rho_\tau\left(\eta_s^*-\frac{\theta' Z_s^A}{\sqrt{n}\sqrt{Sh^k}}\right)-\rho_\tau\left(\eta_s^*\right)|y_s\right]\kappa_s$$
$$=\left(Sh^k\right)^{-1/2}\sum_{s=1}^{S}E\left(\rho_\tau'\left(\eta_s^*\right)|y_s\right)Z_s^{A'}\kappa_s\theta$$
$$+\frac{1}{2}\frac{1}{\sqrt{n}Sh^k}\theta'\sum_{s=1}^{S}E\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)Z_s^A Z_s^{A'}\kappa_s\theta\left(1+o_P\left(1\right)\right).$$

Consider first

$$E\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)=f_\eta\left(\beta_0' y_s^A|y_s\right)\approx\phi\left(\beta_0' y_s^A\middle|\tilde{\eta}\left(y\right),\frac{1}{n}\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y\right)$$
$$=\frac{\sqrt{n}}{\sqrt{\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y}}\phi\left(\frac{\beta_0' y_s^A-\tilde{\eta}\left(y\right)}{\frac{1}{\sqrt{n}}\sqrt{\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y}}\right)$$
$$=\frac{\sqrt{n}}{\sqrt{\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y}}\phi\left(\frac{\sqrt{n}\left(\beta_0' y_s^A-m\left(y_s\right)\right)}{\frac{1}{\sqrt{n}}\sqrt{\partial\eta_y'\mathcal{J}_y^{-1}\partial\eta_y}}+Z_\tau\right).$$

As $y_s\to 0$, $\sqrt{n}\left(\beta_0' y_s^A-m\left(y_s\right)\right)=O_P\left(\sqrt{n}h^{p+1}\right)\to 0$, and

$$\frac{1}{\sqrt{n}}E\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)\to\frac{1}{\sqrt{\partial\eta'\mathcal{J}^{-1}\partial\eta}}\phi\left(Z_\tau\right).$$

Therefore

$$\frac{1}{\sqrt{n}Sh^k}\sum_{s=1}^{S}E\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)Z_s^A Z_s^{A'}\kappa_s\to\frac{1}{\sqrt{n}}\int\frac{1}{h^k}E\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)Z_s^A Z_s^{A'}\kappa_s f\left(y_s\right)dy_s.$$

By a change of variable $y_s=vh$, this converges to, for $C_\kappa=\int v_A v_A'\kappa\left(v\right)dv$,

$$H=C_\kappa\frac{1}{\sqrt{\partial\eta'\mathcal{J}^{-1}\partial\eta}}\phi\left(Z_\tau\right)f_y^\infty\left(0\right).$$

Hence we can write, for θ in a compact set,

$$G_S\left(\theta\right)=\frac{1}{2}\theta' H\theta+\left(Sh^k\right)^{-1/2}\sum_{s=1}^{S}\rho_\tau'\left(\eta_s^*\right)Z_s^{A'}\kappa_s\theta+R_n\left(\theta\right)+o_P\left(1\right).$$

Next we show that $R_S\left(\theta\right)=o_P\left(1\right)$. Since $ER_S\left(\theta\right)=0$, it suffices to bound $ER_S\left(\theta\right)^2$,

$$ER_S\left(\theta\right)^2\leq nSE\left[\left(\rho_\tau\left(\eta_s^*-\frac{\theta' Z_s^A}{\sqrt{n}\sqrt{Sh^k}}\right)-\rho_\tau\left(\eta_s^*\right)\right)-\frac{1}{\sqrt{n}\sqrt{Sh^k}}\rho_\tau'\left(\eta_s^*\right)\theta' Z_s^A\right]^2\kappa_s^2$$
$$\leq nSE\left[1\left(\left|\eta_s^*\right|\leq\frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s^A\right)\left(\frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s^A\right)^2\right]\kappa_s^2$$
$$\leq E\frac{1}{h^k}\left(\theta' Z_s^A\right)^2\kappa_s^2 P\left(\left|\eta_s^*\right|\leq\frac{1}{\sqrt{n}\sqrt{Sh^k}}\theta' Z_s^A|y_s\right)=O\left(\frac{1}{\sqrt{Sh^k}}\right)=o_P\left(1\right),$$

as long as $Sh^k \to \infty$, since $P\left(|\eta_s^*| \leq x/\sqrt{n}|y_s\right) = O\left(x\right)$, and $f_y\left(0\right) = O\left(1\right)$, and that $\theta' Z_s \leq M$ for M the support of κ_s. Next, we show that

$$W_s \equiv \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} \rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s$$

is stochastically bounded and asymptotically normal, so that $\hat{\theta} - H^{-1} W_S = o_p\left(1\right)$, by checking both $Var\left(W_S\right)$ and $E\left(W_S\right)$.

$$\begin{aligned} Var\left(W_S\right) =& \frac{1}{h^k} Var\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s\right) \\ =& \frac{1}{h^k}\left[E_{y_s} Var\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s | y_s\right) + Var_{y_s} E\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s | y_s\right)\right]. \end{aligned}$$

Recall that $\rho_\tau'\left(\eta_s^*\right) = \tau - 1\left(\eta_s - \theta_0' y_s^A \leq 0\right)$, it can be calculated that

$$\begin{aligned} E_{y_s} Var\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) =& E_{y_s} Z_s Z_s' \kappa_s^2 Var\left(\tau - 1\left(\eta_s \leq \theta_0' y_s^A\right) | y_s\right) \\ =& \int Z_s Z_s \kappa_s^2 P\left(\eta_s \leq \theta_0' y_s^A | y_s\right)\left(1 - P\left(\eta_s \leq \theta_0' y_s^A | y_s\right)\right) f\left(y_s\right) dy_s. \end{aligned}$$

By the usual change of variable $y_s = vh$, for $\bar{C}_\kappa = \int v_A v_A' \kappa^2\left(v\right) dv$,

$$\frac{1}{h^k} E_{y_s} Var\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s | y_s\right) \to \tau\left(1 - \tau\right) \bar{C}_\kappa f_y^\infty\left(0\right).$$

Next,

$$Var_{y_s} E\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) = Var_{y_s} Z_s \kappa_s E\left(\tau - 1\left(\eta_s \leq a_0 + b_0 y_s\right) | y_s\right)$$

$$= \left(1 + o(1)\right) Var_{y_s}\left[Z_s \kappa_s f_\eta\left(Q_\tau\left(\eta | y_s\right)\right)\left(o_P\left(\frac{1}{\sqrt{n}}\right) + O\left(h^{p+2}\right) + \sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!} \frac{\partial^{p+1}}{\partial y^u} m\left(0\right)\right)\right].$$

Hence,

$$\frac{1}{h^k n h^{2(p+1)}} Var_{y_s} E\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) \to \tilde{C}_\kappa \frac{1}{\partial \eta' \mathcal{J}^{-1} \partial \eta} \phi^2\left(Z_\tau\right) f_y^\infty\left(0\right),$$

where

$$\tilde{C}_\kappa = \int v_A v_A' \sum_{[u]=p+1} v^u \frac{1}{(p+1)!} \frac{\partial^{p+1}}{\partial y^u} m\left(0\right) \kappa^2\left(v\right) dv.$$

Consider finally the bias term:

$$\begin{aligned} & E \frac{1}{\sqrt{n}\sqrt{Sh^k} h^{p+1}} W_s = E \frac{1}{\sqrt{n} h^k h^{p+1}}\left(\tau - 1\left(\eta_s \leq \theta_0' y_s^A\right) | y_s\right) Z_s \kappa_s \\ =& \frac{1}{h^k h^{p+1}} E Z_s \kappa_s \frac{1}{\sqrt{n}} f_\eta\left(Q_\tau\left(\eta | y_s\right)\right)\left(o_P\left(\frac{1}{\sqrt{n}}\right) + O\left(h^{p+2}\right) + \sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!} \frac{\partial^{p+1}}{\partial y^u} m\left(0\right)\right) \\ \to & \frac{1}{\sqrt{\partial \eta' \mathcal{J}^{-1} \partial \eta}} f_y^\infty\left(0\right) \phi\left(Z_\tau\right) \int v_A \sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!} \frac{\partial^{p+1}}{\partial y^u} m\left(0\right) \kappa\left(v\right) dv. \end{aligned}$$

Hence we conclude when $nh^{2(p+1)} \to 0$, $W_S = O_P\left(1+\sqrt{nSh^k}h^{2(p+1)}\right)$, so that $\hat{\theta} = -H^{-1}W_S + o_P(1) = O_P\left(1+\sqrt{nSh^k}h^{2(p+1)}\right)$. This implies that $\hat{\theta}_{[0]} - \theta_{[0]} = O_p\left(\frac{1}{\sqrt{nSh^k}} + h^{p+1}\right)$. In this case when $Sh^k \to \infty$, simulated posterior quantiles provide a valid inference method as long as the infeasible posterior quantiles $\bar{\eta}_\tau$ are valid.

Similar to the local linear case, the above arguments, strictly speaking, work when $nSh^{2k+2p+2} = O(1)$. In the other case when $nSh^{2k+2p+2} \to \infty$ so that the bias term dominates, a similar sequence of arguments can be followed, but the objective function needs to be normalized differently. In this case, let $\theta = h^{-(p+1)}(b) = h^{-(p+1)}\left(h^{\lfloor u \rfloor}(\beta^u - \beta_0^u, u \in A)\right)$, and redefine

$$\hat{G}_S(\theta) = \frac{1}{h^{2(p+1)}\sqrt{n}Sh^k}\sum_{s=1}^{S}\left(\rho_\tau\left(\eta_s^* - h^{p+1}\theta\right) - \rho_\tau\left(\eta_s^*\right)\right)\kappa_s.$$

Then one can show that in

$$G_S(\theta) = E\left(G_S(\theta)|Y\right) + \frac{1}{h^{p+1}\sqrt{n}Sh^k}\sum_{s=1}^{S}\left(\rho_\tau'\left(\eta_s^*\right)Z_s^{A'}\kappa_s - E\left(\rho_\tau'\left(\eta_s^*\right)|y_s\right)Z_s^{A'}\kappa_s\right)\theta + R_S(\theta),$$

it holds that

$$E\left(G_S(\theta)|Y\right) = \frac{1}{2}\theta' H\theta\left(1+o_P(1)\right) + \frac{1}{h^{p+1}\sqrt{n}Sh^k}\sum_{s=1}^{S}E\left(\rho_\tau'\left(\eta_s^*\right)|y_s\right)Z_s^{A'}\kappa_s\theta.$$

Hence

$$G_S(\theta) = \frac{1}{2}\theta' H\theta + \frac{1}{h^{p+1}\sqrt{n}Sh^k}\sum_{s=1}^{S}\rho_\tau'\left(\eta_s^*\right)Z_s^{A'}\kappa_s\theta + R_n(\theta) + o_P(1).$$

Then in combination of the result that

$$\frac{1}{h^{p+1}\sqrt{n}Sh^k}\sum_{s=1}^{S}\rho_\tau'\left(\eta_s^*\right)Z_s^{A'}\kappa_s \xrightarrow{p} C,$$

for some finite C, it will follow that $\hat{\eta}_\tau - \bar{\eta}_\tau = O_p\left(h^{p+1}\right)$ when $h^{p+1} >> \sqrt{nSh^k}$.

We have therefore completed all the proofs.

References

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

Andrews, D. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2248–2292. North Holland.

——— (1997): "A stopping rule for the computation of generalized method of moments estimators," *Econometrica*, 65(4), 913–931.

Angrist, J. D., and A. B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Belloni, A., and V. Chernozhukov (2009): "On the computational complexity of MCMC-based estimators in large samples," *The Annals of Statistics*, pp. 2011–2055.

Chaudhuri, P. (1991): "Nonparametric estimates of regression quantiles and their local Bahadur representation," *The Annals of Statistics*, 19(2), 760–777.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

Chernozhukov, V., and H. Hong (2003): "A MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115(2), 293–346.

Creel, M. D., and D. Kristensen (2011): "Indirect likelihood inference," .

Fan, J., T.-C. Hu, and Y. K. Truong (1994): "Robust non-parametric function estimation," *Scandinavian Journal of Statistics*, pp. 433–446.

Gallant, A. R., and H. Hong (2007): "A Statistical Inquiry into the Plausibility of Recursive Utility," .

Gallant, R., and G. Tauchen (1996): "Which Moments to Match," *Econometric Theory*, 12, 363–390.

Gentzkow, M., and J. Shapiro (2013): "Measuring the sensitivity of parameter estimates to sample statistics," *Unpublished Manuscript*.

Gourieroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, pp. S85–S118.

Hall, A. R., and A. Inoue (2003): "The large sample behaviour of the generalized method of moments estimator in misspecified models," *Journal of Econometrics*, 114(2), 361–394.

Jun, S. J., J. Pinkse, and Y. Wan (2009): "CUBE–ROOT–N AND FASTER CONVERGENCE, LAPLACE ESTIMATORS, AND UNIFORM INFERENCE," *The Pennsylvania State University working paper*.

——— (2011): "-Consistent robust integration-based estimation," *Journal of Multivariate Analysis*, 102(4), 828–846.

Kim, J., and D. Pollard (1990): "Cube root asymptotics," *Ann. Statist.*, 18(1), 191–219.

Koenker, R., and G. S. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Kristensen, D., and B. Salanié (2010): "Higher order improvements for approximate estimators," *CAM Working Papers*.

Kristensen, D., and Y. Shin (2012): "Estimation of dynamic models with nonparametric simulated maximum likelihood," *Journal of Econometrics*, 167(1), 76–94.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2113–2241. North Holland.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027–1057.

Van der Vaart, A. W., and J. A. Wellner (1996): *Weak convergence and empirical processes*. Springer-Verlag, New York.

Table 1:

Parameter 1				
method	mean bias	median bias	variance	MSE
local constant mean	0.041721	0.027642	0.003294	0.005034
local linear mean	0.030478	0.020029	0.002271	0.003200
local constant median	0.039111	0.025763	0.003083	0.004613
local linear median	0.029184	0.019291	0.002038	0.002889
prior mean	0.064816	0.044197	0.006120	0.010321

Parameter 2				
method	mean bias	median bias	variance	MSE
local constant mean	0.088288	0.060951	0.029003	0.036798
local linear mean	0.070384	0.053991	0.021285	0.026239
local constant median	0.087129	0.060630	0.029493	0.037085
local linear median	0.069867	0.054672	0.023196	0.028077
prior mean	0.103780	0.058788	0.039384	0.050154

Coverage Frequencies		
	Parameter 1	Parameter 2
local constant	0.970	0.910
local linear	0.970	0.890

The number of observations is 100.
The number of simlations is 100.
factor1 is 0.200000.
factor2 is 20.000000.
varscale is 5.000000.

Table 2:

Parameter 1				
method	mean bias	median bias	variance	MSE
local constant mean	0.008508	0.006211	0.000159	0.000231
local linear mean	0.005141	0.004110	0.000103	0.000130
local constant median	0.008200	0.006025	0.000155	0.000222
local linear median	0.004995	0.003859	0.000100	0.000124
prior mean	0.013383	0.011879	0.000198	0.000377

Parameter 2				
method	mean bias	median bias	variance	MSE
local constant mean	0.017261	0.014043	0.001221	0.001519
local linear mean	0.012882	0.011544	0.000821	0.000987
local constant median	0.017080	0.014305	0.001189	0.001481
local linear median	0.012794	0.012221	0.000814	0.000978
prior mean	0.020013	0.014356	0.001739	0.002140

Coverage Frequencies		
	Parameter 1	Parameter 2
local constant	0.980	0.880
local linear	0.920	0.870

The number of observations is 200.
The number of simlations is 100.
factor1 is 0.200000.
factor2 is 40.000000.
varscale is 7.000000.

Table 3:

Parameter 1				
method	mean bias	median bias	variance	MSE
local constant mean	0.014066	0.012260	0.000188	0.000386
local linear mean	0.006917	0.005716	0.000090	0.000138
local constant median	0.013737	0.011657	0.000186	0.000375
local linear median	0.006762	0.005522	0.000088	0.000134
prior mean	0.017001	0.016642	0.000231	0.000520

Parameter 2				
method	mean bias	median bias	variance	MSE
local constant mean	0.024194	0.022537	0.000793	0.001378
local linear mean	0.015911	0.015652	0.000504	0.000757
local constant median	0.024166	0.022921	0.000785	0.001369
local linear median	0.015861	0.015552	0.000505	0.000756
prior mean	0.024926	0.021646	0.000881	0.001503

Coverage Frequencies		
	Parameter 1	Parameter 2
local constant	0.990	0.940
local linear	0.950	0.910

The number of observations is 200.
The number of simlations is 100.
factor1 is 0.400000.
factor2 is 20.000000.
varscale is 7.000000.

Table 4:

Parameter 1				
method	mean bias	median bias	variance	MSE
local constant mean	0.014205	0.013072	0.000170	0.000372
local linear mean	0.005600	0.005348	0.000054	0.000085
local constant median	0.014002	0.012911	0.000169	0.000365
local linear median	0.005652	0.005474	0.000054	0.000086
prior mean	0.016890	0.016324	0.000202	0.000487

Parameter 2				
method	mean bias	median bias	variance	MSE
local constant mean	0.024859	0.021416	0.000763	0.001381
local linear mean	0.013574	0.012447	0.000269	0.000453
local constant median	0.024773	0.021324	0.000748	0.001361
local linear median	0.013443	0.012448	0.000265	0.000445
prior mean	0.026181	0.019455	0.001102	0.001787

Coverage Frequencies		
	Parameter 1	Parameter 2
local constant	1.000	0.920
local linear	0.940	0.910

The number of observations is 400.
The number of simlations is 100.
factor1 is 0.400000.
factor2 is 20.000000.
varscale is 7.000000.

Table 5: Empirical Illustration

parameter	(1)	(2)	(3)	(4)
intercept	4.7758	4.7428	4.7420	4.7690
educ	0.0927	0.0832	0.0835	0.0926
birth year dummies	0.0572	0.0583	0.0583	0.0571
	0.0495	0.0508	0.0508	0.0494
	0.0429	0.0423	0.0423	0.0429
	0.0367	0.0366	0.0366	0.0367
	0.0322	0.0328	0.0328	0.0321
	0.0188	0.0190	0.0190	0.0188
	0.0198	0.0194	0.0194	0.0198
	0.0134	0.0127	0.0127	0.0134
	0.0122	0.0115	0.0114	0.0122

(1) Kernel mean; (2) local linear mean; (3) kernel median; (4) local linear median.

Prior variance = 100 times 2SLS variance

Table 6: Empirical Illustration

parameter	(1)	(2)	(3)	(4)
intercept	5.2063	4.8268	4.8152	5.1934
educ	0.1247	0.1051	0.1055	0.1238
birth year dummies	0.0455	0.0478	0.0480	0.0452
	0.0486	0.0529	0.0527	0.0481
	0.0643	0.0606	0.0600	0.0641
	0.0363	0.0368	0.0364	0.0359
	0.0267	0.0291	0.0288	0.0262
	0.0103	0.0109	0.0109	0.0099
	0.0195	0.0207	0.0213	0.0198
	0.0107	0.0069	0.0071	0.0114
	0.0140	0.0104	0.0102	0.0143

(1) Kernel mean; (2) local linear mean; (3) kernel median; (4) local linear median.

Prior variance = 10000 times 2SLS variance

Table 7: Empirical Illustration

parameter	(1)	(2)	(3)	(4)
intercept	6.2365	5.0240	5.0034	6.2055
educ	0.2012	0.1526	0.1533	0.1975
birth year dummies	0.0172	0.0252	0.0259	0.0164
	0.0463	0.0597	0.0590	0.0451
	0.1158	0.1064	0.1048	0.1150
	0.0353	0.0349	0.0336	0.0342
	0.0136	0.0157	0.0144	0.0120
	-0.0102	-0.0062	-0.0064	-0.0115
	0.0189	0.0234	0.0251	0.0198
	0.0041	-0.0043	-0.0032	0.0063
	0.0184	0.0088	0.0082	0.0191

(1) Kernel mean; (2) local linear mean; (3) kernel median; (4) local linear median.

Prior variance = 100000 times 2SLS variance

Table 8: Empirical Illustration, Serialized version

parameter	(1)	(2)	(3)
intercept	4.7269	4.7229	0.7778
educ	0.0882	0.0817	0.0819
birth year dummies	0.0584	0.0599	0.1390
	0.0495	0.0503	0.0343
	0.0405	0.0412	0.0862
	0.0369	0.0375	0.0770
	0.0328	0.0332	-0.0110
	0.0199	0.0200	-0.0287
	0.0196	0.0195	0.1457
	0.0138	0.0133	0.0618
	0.0119	0.0115	-0.0587

(1) Kernel mean; (2) local linear mean; (3) kernel median.

Prior variance = 100 times 2SLS variance

Table 9: Empirical Illustration, serialized version

parameter	(1)	(2)	(3)
intercept	4.8027	4.7047	12.2387
educ	0.0843	0.0839	-1.0704
birth year dummies	0.0573	0.0580	0.3953
	0.0445	0.0449	-0.6068
	0.0423	0.0442	0.8579
	0.0375	0.0364	0.4346
	0.0334	0.0343	0.5270
	0.0184	0.0181	1.3776
	0.0201	0.0189	-0.6523
	0.0120	0.0109	-0.0965
	0.0114	0.0116	0.8019

(1) Kernel mean; (2) local linear mean; (3) kernel median.

Prior variance = 10000 times 2SLS variance

Table 10: Empirical Illustration, serialized version

parameter	(1)	(2)	(3)
intercept	4.9607	4.6986	28.4655
educ	0.0732	0.0768	-3.5775
birth year dummies	0.0545	0.0547	1.1236
	0.0334	0.0329	-2.0261
	0.0464	0.0510	2.6255
	0.0393	0.0356	1.2949
	0.0349	0.0374	1.5956
	0.0155	0.0152	4.3138
	0.0207	0.0183	-2.1054
	0.0082	0.0054	-0.3347
	0.0102	0.0110	2.5099

(1) Kernel mean; (2) local linear mean; (3) kernel median.

Prior variance = 100000 times 2SLS variance

A Technical Addendum

In this technical addendum we extend several well known results in the literature, namely Theorem 2.1, 7.1 and 7.3 in Newey and McFadden (1994), to allow for their uniform version in $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a shrinking neighborhood of zero. These extensions are used in the intermediate steps in the proof.

First we consider consistency. The following lemma is a straightforward extension of Theorem 2.1 in Newey and McFadden (1994) to allow for uniform convergence in $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a small neighborhood shrinking to zero.

LEMMA 3 Suppose the following three conditions hold. (1) Uniform convergence.

$$\sup_{\theta \in \Theta, y \in \mathcal{Y}} |\hat{Q}(\theta, y) - Q(\theta, y)| = o_P(1);$$

(2) Uniform uniqueness. For all $\epsilon > 0$, there exists $\delta > 0$, such that for any $\tilde{\theta}(\cdot)$ such that $\inf_{y \in \mathcal{Y}} |\tilde{\theta}(y) - \theta(y)| > \delta$, it holds that

$$\sup_{y \in \mathcal{T}} Q\left(\tilde{\theta}(y), y\right) - Q(\theta(y), y) < -\epsilon;$$

(3) For any $\epsilon > 0$, with probability converging to 1, for all $y \in \mathcal{Y}$, $\hat{Q}\left(\tilde{\theta}(y)\right) > \hat{Q}(\theta(y)) - \epsilon$. Then $\sup_{y \in \mathcal{T}} |\tilde{\theta}(y) - \theta(y)| = o_P(1)$.

Proof: Condition (3) is automatically satisfied when $\tilde{\theta}$ is defined as the arg max of $\hat{Q}(\theta, y)$. Its proof directly extends that of Theorem 2.1 in Newey and McFadden (1994). Under the stated conditions (3) and (1), for each $\epsilon > 0$, with probability converging to 1, for all $y \in \mathcal{T}$,

$$Q\left(\tilde{\theta}(y), y\right) > \hat{Q}\left(\tilde{\theta}(y), y\right) - \epsilon/3 > \hat{Q}(\theta(y), y) - 2\epsilon/3 > Q(\theta(y), y) - \epsilon.$$

In the above the first and third inequalities follow from condition (1) and the second inequality follows from condition (3). Finally, given $\delta > 0$, choose $\epsilon > 0$ so that condition (2) holds, then with probability converging to 1, by condition (2),

$$\inf_{t \in \mathcal{T}} Q\left(\tilde{\theta}(y), y\right) - Q(\theta(y), y) > -\epsilon,$$

implies that $\sup_{y \in \mathcal{Y}} |\tilde{\theta}(y) - \theta(y)| < \delta$. ■

Next we generalize Theorem 7.1 in Newey and McFadden (1994) to allow for uniformity in $y \in \mathcal{Y}$. In the following $o_P(\cdot)$ and $O_P(\cdot)$ denote random variables that do not depend on $y \in \mathcal{Y}$ and that satisfy the corresponding stochastic order.

LEMMA 4 Suppose that $\inf_{y\in\mathcal{Y}}\left(\hat{Q}_y\left(\tilde{\theta}_y\right) - \sup_{\theta\in\Theta}\hat{Q}_y(\theta)\right) \geq -o_P(n^{-1})$, and that $\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y - \theta_y| = o_P(1)$. Let assumption 3 hold. Then $\sup_{y\in\mathcal{Y}}|\sqrt{n}\left(\tilde{\theta}_y - \theta_y\right) - J_y^{-1}\sqrt{n}\Delta_n^y| = o_P(1)$.

The proof retraces the steps in Newey and McFadden (1994). First we show that $\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y - \theta_y| = O_P(1)$. By the second part of assumption 3, $\exists C > 0$ such that for all $y \in \mathcal{Y}$ and all $\theta - \theta_y = o(1)$,

$$Q(\theta) - Q(\theta_y) = \frac{1}{2}(\theta - \theta_y)' H_y(\theta - \theta_0) + o_1(1)|\theta - \theta_y|^2 \leq -C|\theta - \theta_y|^2.$$

Since $\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y - \theta_y| = o_P(1)$, with probability converging to 1 (w.p.c.1),

$$Q\left(\tilde{\theta}_y\right) - Q(\theta_y) \leq -C|\tilde{\theta}_y - \theta_y|^2.$$

Note that Assumption 3 also implies that if we had defined

$$\hat{R}^y(\theta, \theta^*) = \hat{Q}_y(\theta) - \hat{Q}_y(\theta^*) - (\theta - \theta^*)'\Delta_n^y - (Q(\theta) - Q(\theta^*))$$

it also holds that for any sequence of $\delta \to 0$

$$\sup_{y\in\mathcal{Y}}\sup_{|\theta-\theta^*|\leq\delta, \theta\in\mathcal{N}(\theta_0), \theta^*\in\mathcal{N}(\theta_0)} \frac{\hat{R}^y(\theta, \theta^*)}{|\theta - \theta^*|^2 + |\theta - \theta^*|/\sqrt{n}} = o_P(1). \tag{9}$$

this implies that w.p.c.1, for all $y \in \mathcal{Y}$,

$$\sqrt{n}R_n^y\left(\tilde{\theta}_y, \theta_y\right)/|\tilde{\theta}_y - \theta_y| \leq \left(1 + \sqrt{n}|\tilde{\theta}_y - \theta_y|\right) o_P(1),$$

so that w.p.c.1, for all $y \in \mathcal{Y}$,

$$\begin{aligned}
0 \leq & \hat{Q}_y\left(\tilde{\theta}_y\right) - \hat{Q}(\theta_y) + o_P\left(n^{-1}\right) = Q_y\left(\tilde{\theta}_y\right) - Q(\theta_y) + \Delta_n^{y'}\left(\tilde{\theta}_y - \theta_y\right) + \hat{R}\left(\tilde{\theta}_y, \theta_y\right) + o_P\left(n^{-1}\right) \\
& \leq -C|\tilde{\theta}_y - \theta_y|^2 + |\tilde{\theta}_y - \theta_y||\Delta_n^{y'}| + |\tilde{\theta}_y - \theta_y|\left(1 + \sqrt{n}|\tilde{\theta}_y - \theta_y|\right) o_P\left(n^{-1/2}\right) + o_P\left(n^{-1}\right) \\
& \leq -(C + o_P(1))|\tilde{\theta}_y - \theta_y|^2 + |\tilde{\theta}_y - \theta_y|\left(\sup_{y\in\mathcal{Y}}\Delta_n^y + o_P\left(n^{-1/2}\right)\right) + o_P\left(n^{-1}\right) \\
& = -\frac{C}{2}|\tilde{\theta}_y - \theta_y|^2 + |\tilde{\theta}_y - \theta_y| O_P\left(n^{-1/2}\right) + o_P\left(n^{-1}\right),
\end{aligned}$$

so that $\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y-\theta_y|\leq O_P\left(n^{-1/2}\right)$ by the same arguments in Newey and McFadden (1994).

Next define $\check{\theta}_y=\theta_y+J_y^{-1}\Delta_n^y$, so that $\sup_{y\in\mathcal{Y}}|\check{\theta}_y-\theta_y|=O_P\left(n^{-1/2}\right)$. By (9), uniformly in $y\in\mathcal{Y}$,

$$\begin{aligned}\hat{Q}\left(\tilde{\theta}_y\right)-\hat{Q}\left(\theta_y\right)=&\frac{1}{2}\left(\tilde{\theta}_y-\theta_y\right)'H_y\left(\tilde{\theta}_y-\theta_y\right)+\Delta_n^{y'}\left(\tilde{\theta}_y-\theta_y\right)+o_P\left(n^{-1}\right)\\=&\frac{1}{2}\left(\tilde{\theta}_y-\theta_y\right)'H_y\left(\tilde{\theta}_y-\theta_y\right)-\Delta_n^{y'}J_y^{-1}H_y\left(\tilde{\theta}_y-\theta_y\right)+o_P\left(n^{-1}\right),\end{aligned}$$

and

$$\begin{aligned}\hat{Q}\left(\check{\theta}_y\right)-\hat{Q}\left(\theta_y\right)=&\frac{1}{2}\left(\check{\theta}_y-\theta_y\right)'H_y\left(\check{\theta}_y-\theta_y\right)+\Delta_n^{y'}\left(\check{\theta}_y-\theta_y\right)+o_P\left(n^{-1}\right)\\=&-\frac{1}{2}\left(\check{\theta}_y-\theta_y\right)'H_y\left(\check{\theta}_y-\theta_y\right)+o_P\left(n^{-1}\right)\end{aligned}$$

Taking difference and noting that uniformly in $y\in\mathcal{Y}$,

$$\hat{Q}\left(\tilde{\theta}_y\right)-\hat{Q}\left(\theta_y\right)-\left(\hat{Q}\left(\check{\theta}_y\right)-\hat{Q}\left(\theta_y\right)\right)\geq o_P\left(n^{-1}\right)$$

it follows that

$$\begin{aligned}o_P\left(n^{-1}\right)\leq&\frac{1}{2}\left(\tilde{\theta}_y-\theta_y\right)'H_y\left(\tilde{\theta}_y-\theta_y\right)-\Delta_n^{y'}J_y^{-1}H_y\left(\tilde{\theta}_y-\theta_y\right)+\frac{1}{2}\left(\check{\theta}_y-\theta_y\right)'H_y\left(\check{\theta}_y-\theta_y\right)\\=&\left(\check{\theta}_y-\theta_y\right)'H_y\left(\check{\theta}_y-\theta_y\right)\leq-C|\tilde{\theta}_y-\check{\theta}_y|^2\end{aligned}$$

Hence conclude that $\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y-\theta_y-J_y^{-1}\Delta_n^y|=\sup_{y\in\mathcal{Y}}|\tilde{\theta}-\check{\theta}_y|=o_P\left(n^{-1/2}\right)$. ■

The next lemma reworks Theorem 7.2 in Newey and McFadden (1994) to verify the GMM model.

LEMMA 5 The conclusions of Lemmas 3 and 4 hold under Assumptions 1 and 4–6.

As in Newey and McFadden (1994) the hypotheses of Theorem 4 are verified. Let $Q_y(\theta)=-\left(g(\theta)-y\right)'W(\theta_y)\left(g(\theta)-y\right)$, where θ_y is defined by $G(\theta_y)'W(\theta_y)\left(g(\theta_y)-y\right)=0$. Uniformly in $y\in\mathcal{Y}$ and $|\theta-\theta_y|=o(1)$, for G_k denoting the kth column of G, write

$$g(\theta)-y=g(\theta_y)-y+G(\theta_y)(\theta-\theta_y)+\sum_{k=1}^{K}\left(\theta^k-\theta_y^k\right)\frac{G_k(\theta_y)}{\partial\theta}(\theta-\theta_y)+o\left(|\theta-\theta_y|^2\right).$$

Hence we can write, uniformly in $y\in\mathcal{Y}$ and $|\theta-\theta_y|=o(1)$,

$$\begin{aligned}-\left(Q_y(\theta)-Q_y(\theta_y)\right)=&2(\theta-\theta_y)'\left(\sum_{i=1}^{d}\left(\sum_{j=1}^{d}W_j^i\left(g_j(\theta_y)-y\right)\right)\frac{\partial^2 g_i(\theta_y)}{\partial\theta\partial\theta'}\right)(\theta-\theta_y)\\&+(\theta-\theta_y)'G(\theta_y)'W(\theta_y)G(\theta_y)(\theta-\theta_y)+o\left(|\theta-\theta_y|^2\right).\end{aligned}$$

This verifies part 2 of Assumption 3, and condition 2 of Lemma 3.

Next we consider the key condition (9). Let $\Delta_n^y = (\hat{g}(\theta_y) - g(\theta_y))' W_y G_y$, for $W_y = W(\theta_y)$ and $G_y = G(\theta_y)$. Also define

$$\hat{\epsilon}(\theta, \theta_y) = \frac{\hat{g}(\theta) - \hat{g}(\theta_y) - g(\theta) + g(\theta_y)}{1 + \sqrt{n}|\theta - \theta_y|}.$$

Then, Assumption 4 implies

$$\hat{\epsilon} \equiv \hat{\epsilon}(\mathcal{Y}, \delta) = \sup_{y \in \mathcal{Y}, |\theta - \theta_y| \leq \delta} \hat{\epsilon}(\theta, \theta_y) = o_P\left(1/\sqrt{n}\right). \tag{10}$$

Recall that $Q_y(\theta) = (g(\theta) - y)' W_y (g(\theta) - y)$, and that $\hat{Q}_y(\theta) = (\hat{g}(\theta) - y)' W_y (\hat{g}(\theta) - y)$. By expanding

$$\hat{g}(\theta) = \hat{g}(\theta_y) + g(\theta) - g(\theta_y) + \epsilon(\theta, \theta_y)\left(1 + \sqrt{n}|\theta - \theta_y|\right).$$

We can decompose

$$R(\theta, \theta_y) = \hat{Q}(\theta) - \hat{Q}(\theta_y) - Q(\theta) + Q(\theta_y) - \Delta_n^{y'}(\theta - \theta_y) = (1) + (2) + (3) + (4) + (5) \tag{11}$$

where we will bound each of the above six terms (in order of magnitude), so that each term is either $o_P(n^{-1})$ or satisfies condition (9).

$$\begin{aligned}
(1) =& 2(g(\theta) - g(\theta_y))' \hat{W}_y (g(\theta) - g(\theta_y)) - Q_y(\theta) + Q_y(\theta_y) \\
=& \underbrace{2(g(\theta) - g(\theta_y))' W_y (g(\theta) - g(\theta_y)) - Q_y(\theta) + Q_y(\theta_y)}_{(1.1)} \\
&+ \underbrace{2(g(\theta) - g(\theta_y))' \left(\hat{W}_y - W_y\right)(g(\theta) - g(\theta_y))}_{(1.2)}.
\end{aligned}$$

Obviously,

$$\frac{\sqrt{n}|(1.2)|}{|\theta - \theta_y|(1 + \sqrt{n}|\theta - \theta_y|)} \leq |\hat{W}_y - W_y| \frac{|g(\theta) - g(\theta_y)|^2}{|\theta - \theta_y|^2} = o_P(1).$$

Noting that

$$G(\theta_y)' W_y (g(\theta_y) - y) = 0,$$

we can write (abusing dimension notations)

$$|(1.1)| = 2\left(g\left(\theta\right) - g\left(\theta_y\right)\right)' W_y \left(g\left(\theta_y\right) - y\right) = |H\left(\theta_y^*\right)||\theta - \theta_y|^2 |W_y||g\left(\theta_y\right) - y|$$

Then

$$\frac{\sqrt{n}|\left(1.1\right)|}{|\theta - \theta_y|\left(1 + \sqrt{n}|\theta - \theta_y|\right)} \leq |H\left(\theta_y^*\right)||W_y||g\left(\theta_y\right) - y| = o\left(1\right)$$

since $y \to 0$. The second term

$$(2) = \left(1 + \sqrt{n}|\theta - \theta_y|\right)^2 \hat{\epsilon}' \hat{W}_y \hat{\epsilon}$$

can be handled in the same way as in Newey and McFadden (1994).

$$(3) = \left(\hat{g}\left(\theta_y\right) - y\right)' \hat{W}_y \left(g\left(\theta\right) - g\left(\theta_y\right)\right) - \Delta_n^{y'}\left(\theta - \theta_y\right) = (3.1) + (3.2).$$

$$(3.1) = \left(\hat{g}\left(\theta_y\right) - g\left(\theta_y\right)\right)' \hat{W}_y \left(g\left(\theta\right) - g\left(\theta_y\right)\right) - \Delta_n^{y'}\left(\theta - \theta_y\right)$$

$$(3.2) = \left(g\left(\theta_y\right) - y\right)' \hat{W}_y \left(g\left(\theta\right) - g\left(\theta_y\right)\right)$$

Consider first $(3.2) = (3.2.1) + (3.2.2)$, where

$$(3.2.1) = \left(g\left(\theta_y\right) - y\right)' W_y \left(g\left(\theta\right) - g\left(\theta_y\right)\right) = (1.1)$$

and

$$(3.2.1) = \left(g\left(\theta_y\right) - y\right)' \left(\hat{W}_y - W_y\right)\left(g\left(\theta\right) - g\left(\theta_y\right)\right)$$

$$\frac{\sqrt{n}|\left(1.1\right)|}{|\theta - \theta_y|\left(1 + \sqrt{n}|\theta - \theta_y|\right)} \leq \sqrt{n}|\hat{W}_y - W_y| \frac{|g\left(\theta\right) - g\left(\theta_y\right)|}{|\theta - \theta_y|} |g\left(\theta_y\right) - y| = o_P\left(1\right). \tag{12}$$

Under the additional condition in Assumption 6.(2)

$$\sup_{y \in \mathcal{Y}} \sqrt{n}|\hat{W}_y - W_y| = O_P\left(1\right), \tag{13}$$

since we also have $|g\left(\theta_y\right) - y| = o\left(1\right)$.

Remark: Even without condition (b) in Assumption 6, (12) still holds for the set of θ such that $|g\left(\theta_y\right) - y| \leq C_y|\theta - \theta_y|$ for each given y, where C_y is bounded away from 0 and ∞.

Note that for exactly identified models, $|g(\theta_y) - y| = 0$ automatically, so this restriction is only binding in overidentified models. For this restricted set of parameters,

$$\frac{\sqrt{n}|(1.1)|}{|\theta-\theta_y|(1+\sqrt{n}|\theta-\theta_y|)} \leq |\hat{W}_y - W_y| \frac{|g(\theta)-g(\theta_y)|}{|\theta-\theta_y|} \frac{|g(\theta_y)-y|}{|\theta-\theta_y|} \leq C|\hat{W}_y - W_y| = o_P(1).$$

Consequently, this implies that it is only necessary to further investigate stochastic equicontinuity for $|\theta - \theta_y| = O_p(|g(\theta_y) - y|)$, which in turn is no larger than the order of the bandwidth in the kernel function. In particular, if the bandwidth $h = O\left(n^{-1/2}\right)$, then $|\theta - \theta_y|$ is automatically $\sqrt{n}$ consistent.

Next write (3.1) = (3.1.1) + (3.1.2),

$$(3.1.2) = (\hat{g}(\theta_y) - g(\theta_y))' \left(\hat{W}_y - W_y\right)(g(\theta) - g(\theta_y)) = O_P\left(\frac{1}{\sqrt{n}}\right) o_P(1) O(|\theta - \theta_y|).$$

$$\begin{aligned}(3.1.1) &= (\hat{g}(\theta_y) - g(\theta_y))' W_y (g(\theta) - g(\theta_y)) - \Delta_n^{y'}(\theta - \theta_y) \\ &= (\hat{g}(\theta_y) - g(\theta_y))' W_y (g(\theta) - g(\theta_y) - G_y(\theta - \theta_y)) = O_P\left(\frac{1}{\sqrt{n}}\right) O\left(|\theta - \theta_y|^2\right).\end{aligned}$$

$$(4) = (g(\theta) - g(\theta_y))' \hat{W}_y \hat{\epsilon} \left(1 + \sqrt{n}|\theta - \theta_y|\right) = O(|\theta - \theta_y|) o_P\left(\frac{1}{\sqrt{n}}\right)\left(1 + \sqrt{n}|\theta - \theta_y|\right)$$

Finally, consider

$$(5) = (\hat{g}(\theta) - y) \hat{W}_y \hat{\epsilon} \left(1 + \sqrt{n}|\theta - \theta_y|\right) = (5.1) + (5.2)$$

$$\begin{aligned}(5.1) &= (\hat{g}(\theta) - g(\theta_y)) \hat{W}_y \hat{\epsilon} \left(1 + \sqrt{n}|\theta - \theta_y|\right) \\ &= O_P\left(\frac{1}{\sqrt{n}}\right) o_P\left(\frac{1}{\sqrt{n}}\right)\left(1 + \sqrt{n}|\theta - \theta_y|\right) = o\left(n^{-1}\right) + o_P\left(\frac{1}{\sqrt{n}}|\theta - \theta_y|\right).\end{aligned}$$

The last term

$$(5.2) = (g(\theta_y) - y)' \hat{W}_y \hat{\epsilon} \left(1 + \sqrt{n}|\theta - \theta_y|\right) = (g(\theta_y) - y)' \hat{W}_y (\hat{g}(\theta) - \hat{g}(\theta_y) - (g(\theta) - g(\theta_y)))$$

seems the most difficulty to deal with. This term is not present when $y = 0$, since $g(\theta_0) = 0$ as long as the model is correctly specified. However, since our approach depends on the local behavior when y is close to but not equal to zero, local misspecification becomes an important part of the analysis.

Case Assumption 7 (a): When the model is exactly identified, then $g(\theta_y) - y = 0$ for all y, and this term (5.2) vanishes.

Case Assumption 7 (b): This case applies when the model is overidentified and when the moment condition is smoothly differentiable. Under this assumption,

$$(\hat{g}(\theta) - \hat{g}(\theta_y) - (g(\theta) - g(\theta_y))) = O_P\left(\frac{1}{\sqrt{n}}\right)|\theta - \theta_y|$$

then we can write

$$\begin{aligned}(5.2) &= (g(\theta_y) - y)' \hat{W}_y O_P\left(\frac{1}{\sqrt{n}}\right)|\theta - \theta_y| \\ &= o(1)\hat{W}_y O_P\left(\frac{1}{\sqrt{n}}|\theta - \theta_y|\right) = o_P\left(\frac{1}{\sqrt{n}}|\theta - \theta_y|\right).\end{aligned}$$

as required.

Case Assumption 7 (c): This is the most difficult case where we need to invoke the additional assumption that $\sup_{y\in\mathcal{Y}}|y| = o\left(n^{-\frac{1}{4}}\right)$. We first show that under this condition, using similar arguments as in Kim and Pollard (1990), that

$$\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y - \theta_y| = o_P\left(\frac{1}{\sqrt{n}}\right). \tag{14}$$

Then it will follow that immediately that $(14) = o_P(n^{-1})$ uniformly over $y \in \mathcal{Y}, |y| = o\left(n^{-\frac{1}{4}}\right), |\theta - \theta^y| \leq \frac{1}{\sqrt{n}}$.

Arguments leading to (14) in the style of Kim and Pollard (1990) are as follows. First note that when $\sup_{y\in\mathcal{Y}}|y| = o\left(n^{-\frac{1}{4}}\right)$,

$$\sup_{y\in\mathcal{Y}}|g(\theta_y) - y| = O\left(\sup_{y\in\mathcal{Y}}|y|\right) = o\left(n^{-1/4}\right).$$

Consider again the decomposition in (11), for some constant $C > 0$,

$$\begin{aligned}\hat{Q}(\theta) - \hat{Q}(\theta_y) =& Q(\theta) - Q(\theta_y) + \Delta_n^{y'}(\theta - \theta_y) + (1) + \ldots + (5) \\ =& -(C + o_P(1))|\theta - \theta_y|^2 + O_P\left(\frac{1}{\sqrt{n}}\right)|\theta - \theta_y| + O_P\left(\frac{\sup_{y\in\mathcal{Y}}|y|}{\sqrt{n}}\right)\sqrt{|\theta - \theta_y|} + o_P\left(n^{-1}\right).\end{aligned}$$

Note that by definition $\hat{Q}\left(\tilde{\theta}_y\right) - \hat{Q}(\theta_y) \geq -o_P(n^{-1})$. This can be used to show by contradiction that $|\tilde{\theta}_y - \theta_y| = O_P\left(n^{-1/2}\right)$.

Suppose not, then for each $\epsilon > 0$, there exists a sequence $M_n^\epsilon \to \infty$, such that

$$\lim_{n\to\infty} \inf P\left(\sqrt{n}|\tilde{\theta}_y - \theta_y| > M_n^\epsilon\right) > \epsilon.$$

Then with strictly positive probability infinitely often,

$$\begin{aligned}\hat{Q}(\theta) - \hat{Q}(\theta_y) &= -(C + o_P(1)) n^{-1} (M_n^\epsilon)^2 + O_P\left(\frac{1}{\sqrt{n}}\right) n^{-\frac{1}{2}} (M_n^\epsilon) + O_P\left(\frac{n^{-\frac{1}{4}}}{\sqrt{n}}\right) n^{-\frac{1}{4}} \sqrt{M_n^\epsilon} + o_P\left(n^{-1}\right) \\ &= -(C + o_P(1)) n^{-1} (M_n^\epsilon)^2 + O_P\left(n^{-1}\right) (M_n^\epsilon) + o_P\left(n^{-1}\right) \sqrt{M_n^\epsilon} + o_P\left(n^{-1}\right)\end{aligned}$$

For sufficiently large M_n^ϵ, this contradicts $\hat{Q}\left(\tilde{\theta}_y\right) - \hat{Q}(\theta_y) \geq -o_P(n^{-1})$.

More concisely, under assumption 7[c], uniformly in $y \in \mathcal{Y}$,

$$R_y(\theta, \theta_y) = o_P\left(\frac{1}{n}\right) + o_P\left(\frac{1}{\sqrt{n}}|\theta - \theta_y|\right) + o_P\left(\frac{1}{\sqrt{n}} \frac{1}{n^{-\frac{1}{4}}} \sqrt{|\theta - \theta_y|}\right) + o_P\left(|\theta - \theta_y|^2\right).$$

Note that

$$\frac{1}{\sqrt{n}} \frac{1}{n^{-\frac{1}{4}}} \sqrt{|\theta - \theta_y|} \leq 2\frac{1}{n} + \frac{1}{\sqrt{n}}|\theta - \theta_y|.$$

Hence it also satisfies the requirement in Theorem 7.1 in Newey and McFadden (1994) that

$$R_y(\theta, \theta_y) = o_P\left(\frac{1}{n}\right) + o_P\left(\frac{1}{\sqrt{n}}|\theta - \theta_y|\right) + o_P\left(|\theta - \theta_y|^2\right).$$

Case assumption 7 (d): When $g(\theta_y) - y = O\left(n^{-1/2}\right)$, it follows immediately from Assumption 4 that $(5.2) = o_P(n^{-1})$. Of course the requirement that $|y| = O\left(n^{-1/2}\right)$ is arguably strong and computationally intensive.

This concludes the proof of Lemma 5. ■

Proof Lemma 1 : This is done by carefully going over the steps in the proof in Chernozhukov and Hong (2003) to show that convergence of each term is uniformly in $y \in \mathcal{Y}$ that shrinks to zero. In particular, the uniform version of the key assumption 4 ((i) and (ii)) in Chernozhukov and Hong (2003) can be written as requring that uniformly in $y \in \mathcal{Y}$ and in $|\theta - \theta_y| = o(1)$,

$$R_y(\theta, \theta_y) = o_P\left(\frac{1}{n} + |\theta - \theta_y|^2\right).$$

Since we have

$$\frac{1}{\sqrt{n}}|\theta-\theta_y| \leq 2\left(\frac{1}{n}+|\theta-\theta_y|^2\right),$$

Assumption 3, which stipulates that $R_y(\theta,\theta_y) = o_P\left(\frac{1}{\sqrt{n}}|\theta-\theta_y|+|\theta-\theta_y|^2\right)$, immmediately implies assumption 4(i) and 4(ii)

Remark: To recap, condition (c) in assumption 7 is applicable to overidentified quantile IV methods or simulated method of moments, in which the moment conditions take for example the form of

$$\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} Z_i\left(1\left(y_i \leq x_i'\theta\right)-\tau\right)$$

and in which the dimension of Z_i is larger than the dimension of θ. Under this condition, using the same arguments in the proof of Lemma 5, it can be shown that for any $\sup_{y\in\mathcal{Y}}|y| \to 0$, $|\tilde{\theta}_y-\theta_y| = o_P\left(n^{-\frac{1}{3}}\right)$. Therefore, for nonsmooth moment conditions for which Assumption 7(c) applies, we suggest an iteration approach. In the first step, one can use a larger $h \to$ in combination with a local polynomial regression of sufficiently high order. This will bring the posterior distribution of θ into a $o_P\left(n^{-1/3}\right)$ neighborhood of the true parameter. In the second step, or subsequent iterative steps, one chooses a smaller $h = o\left(n^{-\frac{1}{4}}\right)$ and sample from the neighborhood of the initial parameter estimate. Using a local linear or local polynomial regression, $\sqrt{n}$ consistency and asymptotic normality will be achieved. It is natural to expect that estimation based on nonsmooth moment conditions should be more difficult and requires more computational efforts.

The theoretical vadility of this iterative procedure can be formally justified by adapting the analysis in Jun, Pinkse, and Wan (2009). For $\sup_{y\in\mathcal{Y}}|y| = o(1)$, the arguments in Theorem 3 in Jun, Pinkse, and Wan (2009) can be extended to show that, uniformly over $y\in\mathcal{Y}$, $\bar{\theta}_y-\theta_y = O_p\left(n^{-1/3}\right)$. In particular, since the scaling of the objective function is by $n >> n^{2/3}$, a uniform in $y\in\mathcal{Y}$ version of result (ii) of Theorem 3 in Jun, Pinkse, and Wan (2009) holds, which also shows that $\bar{\theta}_y-\tilde{\theta}_y = O_P\left(n^{-1/3}\right)$. Therefore for any $h = o(1)$, a local polynomial regression of degree p will produce

$$\hat{\theta}-\theta = O_P\left(n^{-\frac{1}{3}}\left(1+\frac{1}{\sqrt{Sh^k}}\right)+h^{p+1}\right).$$

Under an initial choice of $h = o\left(n^{-\frac{1}{3(p+1)}}\right)$ and $Sh^k \to \infty$, the first step estimator will satisfy $\hat{\theta} - \theta_0 = O_P\left(n^{-1/3}\right)$. Subsequently, the second step can focus on a shrinking neighborhood of the initial estimator, by choosing $h = o\left(n^{-1/4}\right)$. A local linear or polynomial regression in the second step, using simulated parameters centered at the first stage estimator with $h = o\left(n^{-1/4}\right)$ will produce a $\sqrt{n}$ consistent and asymptotically normal estimator $\hat{\theta}$. Similarly, in the second step, local linear or local quantile regressions can also be used to estimate the quantiles of the posterior distribution, which can be used to form asymptotic valid confidence intervals in a frequentist sense.