MONASH
BUSINESS
SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Bayesian Indirect Inference and the ABC of GMM

Michael Creel, Jiti Gao, Han Hong and Dennis Kristensen

February 2016

Working Paper 01/16

Bayesian Indirect Inference and the ABC of GMM *

Michael Creel, Jiti Gao, Han Hong and Dennis Kristensen

February 9, 2016

Abstract

We propose and study local linear and local polynomial based nonparametric regression methods for implementing Approximate Bayesian Computation (ABC) style indirect inference and GMM estimators. These estimators do not need to rely on numerical optimization or Markov Chain Monte Carlo (MCMC) simulations. They provide an effective complement to the classical M-estimators and to MCMC methods, and can be applied to both likelihood and method of moment based models. We provide formal conditions under which frequentist inference is asymptotically valid and demonstrate the validity of estimated posterior quantiles for confidence interval construction. 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 apply to both exactly and over identified models.

Keywords: GMM Estimators, Laplace Transformations, ABC Estimators, Nonparametric Regressions, Simulation-Based Estimation.

JEL Classification: C12, C15, C22, C52.

*Author Affiliations: Michael Creel (Universitat Autònoma de Barcelona, Barcelona GSE, and MOVE); Jiti Gao (Monash University); Han Hong (Stanford University); Dennis Kristensen (University College London, CEMMAP, and CREATES). We thank helpful comments by Victor Chernozhukov, Xiaohong Chen, Ron Gallant, Hide Ichimura, Michael Jansson, Sung Jae Jun, Joris Pinkse, Jim Powell and participants in various conferences and seminars, and Tingting Cheng for able research assistance. The authors acknowledge support from Spanish MEC grant ECO2014-52506-R and Severo Ochoa Programme for Centres of Excellence in RD (SEV-2015-0563); by an Australian Research Council Professorial Fellowship Award: DP1096374 and an Australian Research Council Discovery Projects Scheme under Grant number: DP130104229; financial support by the National Science Foundation (SES 1459975); both the Department of Economics and SIEPR at Stanford.

1 Introduction and Literature

A building block of econometric analysis is the GMM estimator (Hansen (1982)) and its variants based on auxiliary models and indirect inference (Gallant and Tauchen (1996), Gourieroux et al. (1993), Pakes and Pollard (1989)). Despite extensive efforts, see notably Andrews (1997), the difficulty of numerical optimization remains a formidable impediment in the implementation of these methods. Indirect inference models are also closely related to a parallel literature of ABC (approximate Bayesian computation) in statistics.

In this paper we develop computationally attractive Bayesian indirect inference estimators and ABC style GMM estimators that are based on local linear and local polynomial implementations. These methods combine simulation with nonparametric regression in the computation of GMM and Indirect Inference estimators. They only require simulating the model or computing the moment conditions and running a single set of nonparametric least square regressions both for obtaining a point estimate and a valid confidence interval, and are completely amenable to parallel computing on multiple machines. There is no 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 moment based models.

Our paper builds on results from two previous working papers: Creel and Kristensen (2011) (CK) who first proposed simulated Bayesian Indirect Inference estimators in econometrics and Gao and Hong (2014) (GH) who proposed ABC style GMM estimators, and is also closely related to a large ABC literature and to Chernozhukov and Hong (2003). Our key contribution is the development of new theoretical results regarding the implementation of the local linear and local polynomial estimators. In particular, we derive low bounds on the number of simulations in terms of the order of magnitude needed to achieve parametric rates of convergence and asymptotic normality, that can be expressed as functions of the sample size, the number of parameters, and the degree of polynomials. A reduction in the requisite number of simulations can only be achieved by increasing the degree of polynomial extrapolation and not by higher order kernel functions. In particular, higher order local polynomial methods are computationally more efficient because they reduce both variance and bias, while higher order kernels only serve to reduce bias in these regressions. These results hold for both exactly identified and over-identified models. Furthermore, we prove the asymptotic frequentist validity of confidence intervals constructed using simulated quantiles of the quasi-posterior distribution, which are obtained by running two local linear or polynomial quantile

regressions at two relevant quantile levels. These results provide the theoretical background for further development and exploitation of indirect inference and GMM-ABC methods.

To summarize its computational advantage, the method we study only requires the ability to simulate from the model for each parameter value θ to compute a fixed dimensional summary statistics T_n, or to compute the moment conditions, and the ability of run flexible (nonparametric) least square and quantile regressions for both point estimation and confidence interval construction. The estimator is consistent, asymptotically normal, and asymptotically as efficient as a limited information maximum likelihood estimator. It does not require either optimization, or MCMC, or the complex evaluation of the likelihood function.

A closely related paper in the vast statistics ABC literature is Beaumont et al. (2002), who to our knowledge is the first to propose local linear least square regression, but without theoretical justification. We develop a complete asymptotic theory, formalize the validity of simulated posterior inference, and generalize to nonlinear and nonseparable GMM models. Recently, Gentzkow and Shapiro (2014) also 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 directly concerned about the identifying relation between moments and parameters. We also use nonparametric regressions instead of linear regressions. Furthermore, Jun et al. (2015) and Jun et al. (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. Finally, a recent paper by Forneron and Ng (2015) provides a comprehensive framwork incorporating ABC, Indirect Inference and Laplace Estimators, and analyzes their higher order asymptotic bias properties.

In Section 2 below, we describe the estimation and inference methods, starting with the Bayesian indirect inference estimator and proceeding to a generalization to the GMM model. Section 3 illustrates the methods giving results from finite sample simulation studies. Section 4 develops the asymptotic distribution theory. Section 5 provides an illustrative analytical example. Finally, section 6 concludes. In various sections, we also discuss issues related to different simulation sample sizes and misspecification.

2 The setup and estimators

This section presents the estimation context and the proposed estimators. We begin with the Bayesian indirect inference method for parametric models first proposed in Creel and Kristensen

(2011) in section 2.1, and then present the generalization to the GMM framework (Gao and Hong (2014)) in section 2.2, by relating to the Laplace transformation principle in Chernozhukov and Hong (2003). We discuss the relation between these methods, summarize the key theoretical results, and discuss practical implementation details, before presenting examples and a complete asymptotic theory in the following sections.

2.1 Bayesian Indirect Inference

Consider a fully specified model indexed by a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. Given a sample generated at the unknown true parameter value θ_0, indirect inference type methods make use of a set of summary statistics $T_n \in \mathbb{R}^d$ that are functions of the sample. These could be a set of sample moments or some other more complicated sample statistics, such as the parametric estimates from a computationally feasible auxiliary model. For example, the efficient method of moments (Gallant and Tauchen (1996)) defines T_n to be the score vector of an auxiliary model. The statistics T_n define a limited information likelihood function $f_n(T_n|\theta)$, and given a prior density $\pi(\theta)$, a limited information Bayesian posterior distribution:

$$f_n(\theta|T_n) = \frac{f_n(T_n,\theta)}{f_n(T_n)} = \frac{f_n(T_n|\theta)\pi(\theta)}{\int_\Theta f_n(T_n|\theta)\pi(\theta)\,d\theta}. \tag{1}$$

Information from the Bayesian posterior can be used to conduct valid frequentist statistical inference. For example, the posterior mean

$$\bar{\theta} = \int_\Theta \theta f_n(\theta|T_n)\,d\theta \equiv E_n(\theta|T_n) \tag{2}$$

is consistent and asymptotically normal. Posterior quantiles can also be used to form valid confidence intervals under correct model specification. For each $\tau \in (0,1)$, the posterior τth quantile of the jth parameter, defined as $\bar{\theta}^j_\tau$, is given through the relation (assuming continuity of $f_n(\theta|T_n)$):

$$\int^{\bar{\theta}^j_\tau} f_{nj}(\theta_j|T_n)\,d\theta_j = \tau.$$

In the above, $f_{nj}(\theta_j|T_n)$ is the marginal posterior distribution of θ_j given T_n implied by $f_n(\theta|T_n)$:

$$f_{nj}(\theta_j|T_n) = \int f_n(\theta|T_n)\,d\theta_{-j}.$$

Then a valid $1-\tau$ level confidence interval for θ_j is given by $\left(\bar{\theta}^j_{\tau/2}, \bar{\theta}^j_{1-\tau/2}\right)$. More generally, let $\eta(\theta)$ be a known scalar function of the parameters. A point estimate of $\eta_0 \equiv \eta(\theta_0)$ can be computed using the posterior mean:

$$\bar{\eta} = \int_\Theta \eta(\theta) f_n(\theta|T_n)\,d\theta \equiv E_n(\eta(\theta)|T_n). \tag{3}$$

To conduct inference, define $\bar{\eta}_\tau$, the posterior τth quantile of η given T_n, through

$$\int 1\left(\eta\left(\theta\right) \leq \bar{\eta}_\tau\right) f_n\left(\theta|T_n\right) d\theta = \tau. \tag{4}$$

An asymptotically valid frequentist confidence interval of level $1-\tau$ can then be given by $\left(\bar{\eta}_{\tau/2}, \bar{\eta}_{1-\tau/2}\right)$, in the sense that

$$\lim_{n\to\infty} P\left(\eta_0 \in \bar{\eta}_{\tau/2}, \bar{\eta}_{1-\tau/2}\right) = 1-\tau. \tag{5}$$

Direct computation of (2), (3) and (3) requires knowledge of the likelihood $f_n\left(T_n|\theta\right)$ in (1), which is often times not analytically available. Instead, we analyze feasible versions of (3) and (4) based on model simulations and nonparametric local linear and local polynomial regressions, as described in the following algorithm:

1. Draw $\theta^s, s=1,\ldots,S$ independently from $\pi\left(\theta\right)$. Compute $\eta^s = \eta\left(\theta^s\right)$ for $s=1,\ldots,S$.

2. For each draw generate a sample from the model at this parameter value of θ^s, then compute the corresponding statistic $T_n^s = T_n\left(\theta^s\right), s=1,\ldots,S$.

3. For a kernel function $\kappa\left(\cdot\right)$ and a bandwidth sequence h, define $\hat{\eta} = \hat{a}$ in the following local linear regression, which is the intercept term in a weighted least square regression of η^s on $T_n^s - T_n$ with weights $\kappa\left(\frac{T_n^s - T_n}{h}\right)$.

$$\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \left(\eta^s - a - b'\left(T_n^s - T_n\right)\right)^2 \kappa\left(\frac{T_n^s - T_n}{h}\right). \tag{6}$$

4. Similarly, define a feasible version of $\bar{\eta}_\tau$ as $\hat{\eta}_\tau = \hat{a}$ as the intercept term in a local linear quantile regression, or a weighted quantile regression with weights $\kappa\left(\frac{T_n^s - T_n}{h}\right)$:

$$\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \rho_\tau\left(\eta^s - a - b'\left(T_n^s - T_n\right)\right) \kappa\left(\frac{T_n^s - T_n}{h}\right). \tag{7}$$

   In the above $\rho_\tau\left(x\right) = \left(\tau - 1\left(x \leq 0\right)\right) x$ is the *check function* in Koenker and Bassett (1978).

The local linear least square and quantile regressions in steps 3 and 4 above can also be generalized to local polynomial least square and quantile regressions using the notations in Chaudhuri (1991). For this purpose, for $u = (u_1, \ldots, u_d)$, a d-dimensional vector of nonnegative integers, let $[u] = u_1 + \cdots + u_d$. Let A be the set of all d-dimensional vectors u such that $[u] \leq p$ and set $s\left(A\right) = \#\left(A\right)$. Let $\beta = \left(\beta_u\right)_{u\in A}$ be a vector of coefficients of dimension $s\left(A\right)$. Also let $y_s = \left(T_n^s - T_n\right)$, and

$$y_s^A = \left(y_s^u = y_{s,1}^{u_1} \ldots y_{s,d}^{u_d}, u \in A\right)'.$$

Define the pth order polynomial

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

Then we replace steps 3 and 4 by

3' Define $\hat{\eta} = \hat{\beta}_{[0]}$, the 0th element of $\hat{\beta}$, corresponding to $u \equiv 0$, or $u_1 = \ldots = u_d = 0$, for

$$\hat{\beta} = \left(\sum_{s=1}^{S} y_s^A y_s^{A'} \kappa\left(\frac{y^s}{h}\right) \right)^{-1} \left(\sum_{s=1}^{S} y_s^A \eta^s \kappa\left(\frac{y^s}{h}\right) \right). \tag{8}$$

4' Define $\hat{\eta}_\tau = \hat{\beta}_{[0]}$, the 0-th element of $\hat{\beta}$, for

$$\hat{\beta} \equiv \arg\min_{\beta} \sum_{s=1}^{S} \rho_\tau\left(\eta^s - \beta' y_s^A\right) \kappa\left(\frac{y_s}{h}\right). \tag{9}$$

In particular, $\hat{\theta}$ and $\hat{\theta}_\tau$ correspond to a vector of $\eta_j(\theta) = \theta_j, j = 1, \ldots, k$. Local linear regression is a special case of local polynomial regression when $p = 1$. It will be shown that under suitable regularity conditions, $\hat{\eta}$ and $\hat{\eta}_\tau$ are consistent if $h \to 0$ and $S \to \infty$ when $n \to \infty$. In order for $\hat{\eta}$ to be first order equivalent to (limited information) MLE and for (5) to hold, we require $\sqrt{n} h^{1+p} \to 0$ and $S h^k \to \infty$, which entails $S / \left(n^{\frac{k}{2(p+1)}} \right) \to \infty$, namely that S is much larger than $n^{\frac{k}{2(p+1)}}$. In particular, as standard in nonparametric regression the bias in $\hat{\theta}$ is of $O(h^p)$. However, the variance is of order $O\left(\frac{1}{nSh^k}\right)$, which is much smaller than that in usual nonparametric regression models. In a local linear regression with $p = 1$, this requires S to be larger than $n^{k/4}$, where $k = \dim(\theta)$. This conditions holds regardless of whether $d = k$ or $d > k$.

Simple sampling from a predetermined prior might not be computationally efficient when many draws of the parameter θ^s lead to simulated summary statistics T_n^s that are far away from the observed T_n, so that the associated parameter draws will have little or no weight in the nonparametric regression. A remedy is to choose the prior $\pi(\theta)$ iteratively or adaptively, so that it becomes dependent on the data, which can be denoted as $\pi(\theta | T_n)$. For example, given a consistent initial estimate $\hat{\theta}_0$ that converges at rate n^{r_1}, $\pi(\theta | T_n)$ can be chosen to be normal with mean $\hat{\theta}_0$ with variance n^{-2r_2}, for both $r_1 \geq r_2 \to \infty$. This can also be implemented through the importance sampling ABC algorithm in Creel and Kristensen (2015). Define importance sampling weights $\omega_s = \pi(\theta^s | T_n) / \pi(\theta^s)$, and replace steps 3 and 4 by

3" $\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \left(\eta^s - a - b'\left(T_n^s - T_n\right)\right)^2 \omega_s \kappa\left(\frac{T_n^s - T_n}{h}\right)$,

4" $\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \rho\left(\eta^s - a - b'\left(T_n^s - T_n\right)\right) \omega_s \kappa\left(\frac{T_n^s - T_n}{h}\right)$.

When we draw θ^s directly from $\pi(\theta | T_n)$ instead of from $\pi(\theta)$, we set $\omega_s \equiv 1$.

2.2 The ABC of GMM

The Bayesian indirect inference method is closely related to the ABC literature. In this section we show that the ABC method can be generalized to the GMM context, with possibly nonlinear and nonseparable moment conditions, in which a complete data generating process need not be fully specified.

The GMM estimator is based on a set of d-dimensional sample and parameter dependent moment conditions $\hat{g}(\theta)$ such that $\hat{g}(\theta) \xrightarrow{p} g(\theta)$ and such that $g(\theta) = 0$ if and only if $\theta = \theta_0$. Often times $\hat{g}(\theta)$ takes the form of a sample average although this need not be the case: $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i, \theta)$, where Z_i is the sample data for the ith observation, and by LLN $g(\theta) = Eg(Z_i; \theta)$. Typically, $\sqrt{n}(\hat{g}(\theta) - g(\theta)) \xrightarrow{d} N(0, \Sigma(\theta))$ and a consistent estimate $\hat{\Sigma}(\theta)$ of $\Sigma(\theta)$ is available. For these models, Chernozhukov and Hong (2003) suggest applying MCMC to the quasi-Bayes posterior density

$$f_n(\theta|\text{GMM}) = \frac{\pi(\theta)\exp\left(n\hat{Q}_n(\theta)\right)}{\int \pi(\theta)\exp\left(n\hat{Q}_n(\theta)\right)d\theta}, \tag{10}$$

where $\hat{Q}_n(\theta) = -\frac{1}{2}\hat{g}(\theta)'\hat{W}(\theta)\hat{g}(\theta)$, and where $\hat{W}(\theta)$ is a possibly data and parameter dependent weighting matrix. An optimal choice is $\hat{W}(\theta) = \hat{\Sigma}(\theta)^{-1}$. Then we can redefine (2), (3) and (4) by replacing $f_n(\theta|T_n)$ with $f_n(\theta|\text{GMM})$:

$$\bar{\theta} = \int_\Theta \theta f_n(\theta|\text{GMM})\,d\theta, \quad \bar{\eta} = \int_\Theta \eta(\theta) f_n(\theta|\text{GMM})\,d\theta, \quad \int 1(\eta(\theta) \le \bar{\eta}_\tau) f_n(\theta|\text{GMM})\,d\theta = \tau. \tag{11}$$

To motivate the quasi-Bayes construction in (10), consider the following statistical experiment: Given θ and the data, draw Y_n from a d-dimensional multivariate normal distribution mean vector $\hat{g}(\theta)$ and variance-covariance matrix $\frac{1}{n}\hat{W}(\theta)^{-1}$. For example, if $\hat{W}(\theta) = \hat{\Sigma}(\theta)^{-1}$, then $Y|\theta \sim N\left(\hat{g}(\theta), \frac{1}{n}\hat{\Sigma}(\theta)\right)$. But other choices of $\hat{W}(\theta)$ can be used, e.g. $\hat{W}(\theta) = I$ or $\hat{W}(\theta) = \hat{\Sigma}\left(\hat{\theta}_0\right)^{-1}$ for a consistent initial estimate $\hat{\theta}_0$. Given that θ is drawn from the prior density $\pi(\theta)$, the posterior density of θ given $Y_m = y$ can be written as

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

Notice that (10) is essentially $f_n(\theta|Y_m = 0)$, if we replace $\pi(\theta)$ in (10) by $\pi(\theta)\det\left(\hat{\Sigma}(\theta)\right)^{-\frac{1}{2}}$, or if $\hat{\Sigma}(\theta)$ does not depend on θ. Therefore we replace (11) by

$$\bar{\theta} = \int_\Theta \theta f_n(\theta|Y_n = 0)\,d\theta, \quad \bar{\eta} = \int_\Theta \eta(\theta) f_n(\theta|Y_n = 0)\,d\theta, \quad \int 1(\eta(\theta) \le \bar{\eta}_\tau) f_n(\theta|Y_n = 0)\,d\theta = \tau. \tag{12}$$

Similar to (2), (3) and (4), (12) are theoretically *infeasible* constructs, but they can be implemented by the following simulation and nonparametric regression algorithm.

1. Draw $\theta^s, s = 1, \ldots, S$ from $\pi(\theta)$. For each θ^s, compute $\hat{g}(\theta^s)$.

2. Draw y_n^s from $Y_n \sim N\left(\hat{g}(\theta^s), \frac{1}{n}\hat{W}(\theta^s)^{-1}\right)$. For $\xi \sim N(0, I_d)$:

$$y_n^s = \hat{g}(\theta^s) + \frac{1}{\sqrt{n}}\hat{W}(\theta^s)^{-1/2}\xi. \tag{13}$$

3. Define $\hat{\eta} = \hat{a}$ in the following local (to zero) linear least square regression:

$$\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \left(\eta^s - a - b'(y_n^s)\right)^2 \kappa\left(\frac{y_n^s}{h}\right). \tag{14}$$

4. Define $\hat{\eta} = \hat{a}$ in the following local (to zero) linear quantile regression:

$$\left(\hat{a}, \hat{b}\right) \equiv \arg\min_{a,b} \sum_{s=1}^{S} \rho_\tau\left(\eta^s - a - b'(y_n^s)\right) \kappa\left(\frac{y_n^s}{h}\right). \tag{15}$$

Similarly to section 2.1, a local polynomial extension can be implemented exactly as in (8) and (9). Results regarding S, h in relation to n in section 2.1, and the possible use of importance sampling, also apply to ABC-GMM. Similar to MCMC, ABC-GMM can be particularly useful with nonsmooth moments, such as those in crude frequency-based simulated method of moment models (Pakes and Pollard (1989)).

2.3 Discussion

BIL and ABC-GMM The ABC-GMM model Bayesian indirect inference estimators are closely related through $y_n^s = T_n^s - T_n$. When the *binding function* $t(\theta)$: $T_n \xrightarrow{p,\theta} t(\theta)$ is known, the moment condition $\hat{g}(\theta)$ can be formed by $T_n - t(\theta)$ and ABC-GMM can be applied. When $t(\theta^s)$ is not analytically known, the proposal in Creel and Kristensen (2011) replaces it with a simulated version y_n^s from θ^s and uses $y_n^s = T_n - T_n^s$. This is tantamount to drawing y_n^s from

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

where ξ_n^s is approximately (but not exactly) a standard normal random vector:

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

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, since it avoids the need to estimate $\hat{\Sigma}(\theta^s)$ in a continuously updating GMM or two step optimal GMM setting.

In ABC-GMM, the optimal choice of $\hat{W}(\theta^s)$ should satisfy $\hat{W}(\theta) - \Sigma(\theta_0) \xrightarrow{p} 0$ when $\theta \xrightarrow{p} \theta_0$. This can be implemented through continuously updating, where $\hat{W}(\theta^s) = \hat{\Sigma}(\theta^s)^{-1}$, or through a two step optimal weighing matrix setup, where $\hat{W}(\theta^s) = \hat{\Sigma}\left(\hat{\theta}_0\right)^{-1}$ and $\hat{\theta}_0$ is an initial $\sqrt{n}$ consistent estimator. An ad hoc choice such as $\hat{W}(\theta^s) = I_d$ still produces $\sqrt{n}$ consistent and asymptotically normal $\hat{\theta}$, $\hat{\eta}$ and $\hat{\eta}_{1/2}$. However, the posterior interval $\left(\hat{\eta}_{\tau/2}, \hat{\eta}_{1-\tau/2}\right)$ no longer forms an asymptotically valid $1-\tau$ confidence interval.

Different simulation sample size The simulation sample size can also differ from the observed sample. T_n^s can be replaced by T_m^s, where possibly $m \neq n$. In step 2 of ABC-GMM, y_n^s can be replaced by

$$y_m^s = \hat{g}(\theta^s) + \frac{1}{\sqrt{m}} W(\theta^s)^{-1/2} \xi \sim N\left(\hat{g}(\theta^s), \frac{1}{m} W(\theta^s)^{-1}\right).$$

It can be shown that when $m \to \infty$, $\hat{\rho} - \rho_0 = O_P\left(\frac{1}{\sqrt{\max(n,m)}}\right)$ and that an asymptotically valid (however conservatively so when $m < n$) $1-\tau$th confidence interval for ρ_0 is given by

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

Only when $m = n$, this confidence interval specializes to $\left(\hat{\eta}_{\tau/2}, \hat{\eta}_{1-\tau/2}\right)$. In the rest of the paper we focus on $m = n$, since $m < n$ does not seem to bring computational efficiency unless the cost of simulation increases with the simulation sample size, and $m > n$ does not increase first order asymptotic efficiency.

Heuristically, we may take $m = \infty$, so that $y_\infty^s = \hat{g}(\theta^s)$ or $y_\infty^s = T_n - t(\theta)$. This can be shown to work fine with exactly identified models in which $d = k$, but may lead to difficulties in overidentified models. When $d > k$, conditional on a realization of θ^s, the event that $\hat{g}(\theta^s) = t$ is not possible for almost all values of t. 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. Furthermore, while local constant kernel methods can still be implemented (however with slower rates of convergence), 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. 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.

Prior distribution sampling Since ABC is a local search method, the effective support of $\pi(\theta)$ (or $\pi(\theta|T_n)$) is more important than its shape. In particular, the true parameter needs to be in the interior of the support for the asymptotic theory to be valid, in which case the first order asymptotic theory is not sensitive to the choice of $\pi(\theta)$. In the absence of real prior information, it is common for researchers to specify the initial $\pi(\theta)$ as a uniform distribution over the Cartesian product of compact intervals of the components of the parameter space. When $\pi(\theta)$ is specified as a Cartesian product of uniform distributions, in addition to using pseudo-number generators to obtain draws from $\pi(\theta)$, it is also possible to adopt quasi-(or sub-) random sequences. For example, Press et al. (1992) suggests a deterministic quadrature scheme that samples each Cartesian grid exactly once. This scheme amounts to combining an equally spaced grid search method with the polynomial extrapolation that we suggested. Alternatively, Press et al. (1992) also discuss using Halton's sequence which is closely related to Cartesian grid search, or more sophisticated Sobol-Antonov-Saleev sequences for sampling from the uniform prior $\pi(\theta)$.

Nearest neighborhood implementation One possible method to choose the window width parameter h is using the nearest neighborhood to 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.

3 Monte Carlo Simulations

This section presents examples of use of Bayesian indirect inference and ABC-GMM estimators, using Monte Carlo simulations. It shows that the estimators can give reliable results in relatively complicated estimation contexts, and it serves to clarify the details of how the estimators may be implemented.

3.1 DSGE Model

First, we use Bayesian indirect inference for estimation of a simple nonlinear DSGE model. Full likelihood-based estimation and inference for such models is complicated by unobserved state variables, which necessitate use of nonlinear filtering methods (Fernández-Villaverde and Rubio-Ramírez (2005); An and Schorfheide (2007)). Also, models may contain fewer shocks than state variables, which can lead to stochastic singularities in linearized models. With estimation by ABC, there is no need for filtering, and nonlinear models may be estimated directly, without linearization or ad hoc addition of measurement errors. Our approach is related to that of Ruge-Murcia (2012), who employs the simulated method of moments (SMM) for the estimation of a DSGE model which is similar to that we describe below. Recall that SMM requires numerical optimization, which can be computationally demanding when the parameter space is large. In a simulation study, Ruge-Murcia (2012) treats a number of the parameters as known, while here we estimate all of the model's parameters.

The model that we consider is as follows: A single good can be consumed or used for investment, and a single competitive firm maximizes profits. The variables are: y output; c consumption; k capital; i investment; n labor; w real wages; and r return to capital. The household maximizes expected discounted utility $E_t \sum_{s=0}^{\infty} \beta^s \left(\frac{c_{t+s}^{1-\gamma}}{1-\gamma} + (1 - n_{t+s})\eta_t \psi \right)$ subject to the budget constraint $c_t + i_t = r_t k_t + w_t n_t$ and the accumulation of capital $k_{t+1} = i_t + (1 - \delta k_t)$. There is a preference shock, η_t, that affects the desirability of leisure. The shock evolves according to

$$\ln \eta_t = \rho_\eta \ln \eta_{t-1} + \sigma_\eta \epsilon_t. \tag{16}$$

The competitive firm produces the good y_t using the technology $y_t = k_t^\alpha n_t^{1-\alpha} z_t$. Technology

shocks z_t also follow an AR(1) process in logarithms: $\ln z_t = \rho_z \ln z_{t-1} + \sigma_z u_t$. The innovations to the preference and technology shocks, ϵ_t and u_t, are independent i.i.d. standard normal random variables. The good y_t can be allocated by the consumer to consumption or investment: $y_t = c_t + i_t$. The consumer provides capital and labor to the firm, and is paid at the rates r_t and w_t, respectively.

Following Ruge-Murcia (2012), we estimate steady state hours, $\bar{n}$, along with the other parameters, excepting ψ, because it is comparatively easy to set priors on $\bar{n}$. Then ψ can be recovered using the estimates of the other parameters. The true parameters values are given in the fourth column of Table 1. True steady state hours, $\bar{n}$, is set to 1/3 of the time endowment. The other parameters are set to values that are intended to be representative of the DSGE literature. Our pseudo-prior is chosen as a uniform distribution over the hypercube defined by the bounds of the parameter space, which are found in columns 2 and 3 of Table 1. The chosen limits cause the pseudo-prior means to be biased for the true parameter values (see Table 1, column 5). The chosen limits are intended to be broad, so that the prior mean is quite uninformative as an estimator of the true parameter values (see Table 1, column 6). The DSGE literature sometimes makes use of fairly strongly informative priors, or fixes certain parameters (e.g., Ruge-Murcia (2012)). Our intention here is to try to estimate all parameters of the model, using biased and weakly informative priors, to show that the estimation procedure is able to extract information about all parameters from the sample data.

[Table 1 about here.]

The model is solved and simulated using Dynare (http://www.dynare.org), using a third order perturbation about the steady state. We assume, in line with much empirical work (see Guerrón-Quintana (2010) for discussion), that all variables except the capital stock are observed and available to use in the computation of statistics. The candidate auxiliary statistics include variable sample means, means of ratios of variables, standard deviations, coefficients of first order autoregressions for each variable in turn, and sample variances and covariances, across equations, of the residuals of first order autoregressions. The first order conditions of the model also suggest some statistics that may be informative. For example, the model implies that $w = \psi \eta c^{\gamma}$, so

$$\log w = \log \psi + \gamma \log c + \log \eta, \tag{17}$$

where the preference shock $\log \eta$ follows an AR(1) process (see eq. 16). Because w and c are observable, equation 17 can be estimated, and the residuals of the estimated model may be used

to construct estimators that should be informative for ρ_η and σ_η. In total, the set of candidate statistics has 40 elements. The statistics chosen for the final estimation were selected using the cross validation procedure of Creel and Kristensen (2015). The final set of selected statistics has 22 elements, and is summarized in Table 2.

[Table 2 about here.]

Given the selected statistics, the ABC estimator is computed using the adaptive importance sampling methods described in Algorithms 2 and 3 of Creel and Kristensen (2015). The importance sampling distribution is generated separately for each Monte Carlo replication. Once the importance sampling distribution is generated, 5000 draws from the importance sampling distribution are made, to perform the final nonparametric fitting step.

The final nonparametric fitting step requires setting the bandwidths of the nonparametric fitting and quantile estimation procedures. We present two sets of results. The first results use bandwidths which were selected experimentally, separately for each parameter, to minimize out of sample RMSE and to optimize 90% confidence interval coverage, over 100 "true" parameter values which were drawn randomly from the prior. This is an entirely feasible procedure, which makes use of only pre-sample information. Then these bandwidths were used to do the nonparametric fitting and quantile estimation, using the 1000 Monte Carlo draws for the true parameter values given in Table 1. Software to perform all of these steps, and to replicate the Monte Carlo results reported here, is available at https://github.com/mcreel/ABCDSGE.

Table 3 gives the ABC estimation results for the 1000 Monte Carlo replications. We report results using local constant, local linear, and local quadratic (omitting cross products) nonparametric fits for the posterior mean. Results using the estimated posterior median are very similar, and are therefore not reported here. The table also gives the proportion of times that the true parameter values lie within the estimated 90% confidence interval, based upon nonparametric estimation of the 0.05 and 0.95 conditional quantiles, using a local constant nonparametric quantile estimator. We see that all versions of the ABC estimator reduce bias and RMSE considerably, compared to the prior biases and RMSEs given in Table 1. The local linear and local quadratic versions perform considerably better, overall, than does the local constant version. The magnitude of the biases of the local linear and local quadratic versions is small, compared to the true parameter values, in column 4 of Table 1. Between the local linear and local constant versions, performance is very similar, except that the local quadratic version has a bit less bias for several parameters. With regard

to confidence interval accuracy, we see, in the 8th column of Table 3, that it is problematic. For the parameters σ_η and $\bar{n}$, confidence intervals are too narrow, on average, while for the parameters β, δ, γ, and ρ_η, they are too broad.

[Table 3 about here.]

The results in Table 3 are based upon bandwidths that use no local information, as they were tuned using draws from the prior, which is biased and quite dispersed, given the true parameter values. In actual practice, one would prefer to use bandwidths that are tuned locally for the realized value of the statistic. One means of doing this is to do estimation exactly as was done to generate the results reported in Table 3, but then, given the realized estimate of the parameters, implement the experimental bandwidth tuning procedure using samples drawn at the parameter estimate, rather than draws from the prior. This would provide a feasible, local, bandwidth tuning procedure. Unfortunately, such a procedure is too costly to implement within a Monte Carlo framework, though it is perfectly feasible when performing a single estimation for a real sample. As an approximation, we instead randomly draw 100 "true" parameter values from the 1000 Monte Carlo realized estimates from the first round, and implement the bandwidth tuning method using these. This gives a fixed set of bandwidths to use for each of a new set of Monte Carlo replications, rather that specific bandwidths for each Monte Carlo replication, which would be more desirable, but which is too costly to implement in the Monte Carlo context. Table 4 gives results for 1000 additional Monte Carlo replications, using bandwidths tuned in this way. We see that bias and RMSE are essentially the same as in Table 3, but that confidence interval coverage is considerably improved, overall, though still somewhat problematic for the parameters β, ρ_η and σ_η.

We also estimated true optimal bandwidths, by implementing the tuning procedure using 100 random samples generated at the true parameter values. When such bandwidths are used, 90% confidence interval coverage is correct, within expected error bounds, for all parameters. This procedure is of course not feasible outside of the Monte Carlo context, but it does confirm the theoretical result that confidence intervals have asymptotically correct coverage, and it lends support to performing local bandwidth tuning by drawing random samples at the first round ABC estimate, as this first round estimate is a consistent estimator of the true parameter.

[Table 4 about here.]

Given the simplicity and good performance of the ABC estimator, we believe that it provides an

interesting alternative to the considerably more complex and computationally demanding methodology of MCMC combined with particle filtering, which can probably be described as the current state of the art for estimation of DSGE models. The practicality of estimation of a complex model using ABC is illustrated by the fact that we have been able to perform bandwidth tuning and 2000 Monte Carlo replications of estimation of this simple but still nonlinear DSGE model, using a single 32 core computer, in less than 72 hours. Once statistics and bandwidths have been selected (which are steps which can be performed before the sample data is available), it takes less than two minutes to perform a single estimation of the model. This final estimation step involves embarrassingly parallel computations (simulation and nonparametric regression), which means that ABC estimation as we have implemented it can be used for estimation of complex models in near real time.

3.2 Quantile IV model

In this section we illustrate the ABC-GMM estimator by applying it to the quantile instrumental variable model of Chernozhukov and Hansen (2005), which uses moment conditions that are not separable between the parameters and the data. For the model $y_i = x_i'\beta + \epsilon_i$, where $Q_\tau(\epsilon_i|z_i) = 0$, we consider the following data generating process: $\epsilon_i = \exp\left(\left(z_i'\alpha\right)^2 \upsilon_i\right) - 1$, where υ_i is such that $Q_\tau(\upsilon_i|z_i) = 0$. In particular, we choose $x_i = (1, \tilde{x}_i)$, where $\tilde{x}_i = \xi_{i1} + \xi_{i2}$, and $z_i = (1, \tilde{z}_{i1}, \tilde{z}_{i2})$, where $\tilde{z}_{i1} = \xi_{i2} + \xi_{i3}$, $\tilde{z}_{i2} = \xi_{i1} + \xi_{i4}$, and the four ξ_{ji} are i.i.d. N(0,1), $j = 1, 2, .., 4$. Thus, the regressor $\tilde{x}_i$ is correlated with each of the two instruments $\tilde{z}_{i1}$ and $\tilde{z}_{i2}$, and the instruments are correlated with one another. Also, $\upsilon_i \sim N(0, 1)$, $\forall i$.

Input parameters for the simulation are α (3×1) and β (2×1). The three alpha parameters, which affect the variance of the error of the regression, are all set to 1/5. The parameter of interest is β, estimation of which is based on the moment condition $\hat{g}(\beta) = \frac{1}{n}\sum_{i=1}^n z_i \left(\tau - 1(y_i \leq x_i'\beta)\right)$. For these moment conditions, the optimal weight matrix does not depend on the parameters, and is the inverse of $\frac{1}{n}\sum_{i=1}^n z_i z_i'$. We set the true values of the parameters to $\beta = (1, 1)$, and the sample size to $n = 200$. The prior is set to a uniform distribution over $(0, 3) \times (0, 3)$, so that the prior mean is biased for the true parameter values. Finally, we set $\tau = 0.5$. We implement the ABC-GMM estimator using the same adaptive importance sampling and bandwidth tuning methods as were described in the section giving the DSGE results. After construction of the importance sampling density, $S = 10000$ simulation draws are used for the final nonparametric estimations of conditional means and quantiles. This procedure is repeated for each of 1000 Monte

Carlo replications. Code (using the Julia language) to replicate the results of this section is at https://github.com/mcreel/QuantileIV.jl. A set of 1000 replications takes approximately 8 minutes to complete, using a computer with 32 computational cores.

Table 5 presents the results, for local constant and local linear versions of the ABC-GMM estimator, using bandwidths that were tuned with draws from the prior. For comparison, we also give results for the prior mean as an estimator, and for the simple instrumental variables estimator. We see that the ABC-GMM estimator is much less biased than the prior and the IV estimators, and that RMSE is also considerably lower. The local linear version is somewhat less biased, and with better precision, than the local constant version. Confidence intervals have quite good coverage. Table 6 gives results using the local tuning procedure, as described above. Bias and RMSE of the ABC-GMM estimator change little, but confidence interval coverage is improved, on average, and is quite reliable, overall.

[Table 5 about here.]

[Table 6 about here.]

4 Asymptotic Distribution Theory

In this section we formalize the assumptions that are needed for the asymptotic validity of the estimators and the confidence intervals, and provide conditions on the order of magnitude of the number of simulations in relation to the sample size for $\sqrt{n}$ consistency and asymptotic normality. Part of the assumptions are related to the *infeasible* estimators and intervals, $\bar{\theta}$, $\bar{\eta}$ and $\left(\bar{\eta}_{\tau/2}, \bar{\eta}_{1-\tau/2}\right)$. They mirror the general results in Chernozhukov and Hong (2003) and Creel and Kristensen (2011). Additional assumptions relate to the *feasible* simulation based estimators and intervals, $\hat{\eta}$, $\hat{\eta}_\tau$, and $\left(\hat{\eta}_{1-\tau/2}, \hat{\eta}_{1-\tau/2}\right)$.

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.

The following assumptions, it is understood that $g(\theta) = t(\theta) - t(\theta_0)$ in the IL model, and similarly $\hat{g}(\theta) = T_n - t(\theta_0)$. Let $Q(y|\theta) = -\frac{1}{2}(g(\theta) - y)' W(\theta)(g(\theta) - y)$, where $W(\theta) = \Sigma(\theta)^{-1}$ in IL or optimally weighted GMM. Also define $\theta(y) = \arg\max_{\theta\in\Theta} Q(y|\theta)$. This is the population

limit of the sample objective functions $\hat{Q}_n(y|\theta) = \frac{1}{n}\log f(T_n + y|\theta)$ or

$$\hat{Q}_n(y|\theta) = -\frac{1}{2}(\hat{g}(\theta) - y)' \hat{W}(\theta)(\hat{g}(\theta) - y). \tag{18}$$

Define $G(\theta) = \frac{\partial}{\partial\theta} g(\theta)$. Also denote $H(\theta) = \frac{\partial}{\partial\theta}\text{vech}W(\theta)$.

ASSUMPTION 2 (1) $g(\theta) = 0$ if and only if $\theta = \theta_0$; (2) $W(\theta)$ is uniformly positive definite and finite on $\theta \in \Theta$; (3) $\sup_{\theta\in\Theta} |\hat{W}(\theta) - W(\theta)| = o_P(1)$; (3) $\sup_{\theta\in\Theta} |\hat{g}(\theta) - g(\theta)| = o_P(1)$; (4) $\{\sqrt{n}(\hat{g}(\theta) - g(\theta)); \theta \in \Theta\} \rightsquigarrow \mathcal{G}_g(\cdot)$, a mean zero Gaussian process with marginal variance $\Sigma(\theta)$; (5) $g(\theta)$ and $W(\theta)$ are both $p+1$ times continuously differentiable with bounded derivatives. (6) For any $\epsilon > 0$, there is $\delta > 0$ such that

$$\limsup_{n\to\infty} P\left\{\sup_{|\theta-\theta'|\leq\delta} \frac{\sqrt{n}|(\hat{g}(\theta) - \hat{g}(\theta')) - (g(\theta) - g(\theta'))|}{1 + \sqrt{n}|\theta - \theta'|} > \epsilon\right\} < \epsilon. \tag{19}$$

ASSUMPTION 3 The model is exactly identified: $d = k$.

ASSUMPTION 4 There exists random functions $\hat{G}(\theta_y)$ and $\hat{H}(\theta_y)$, such that for any $\delta_n \to 0$,

$$\sup_{|\theta-\theta_y|\leq\delta_n} \sup_{y\in\mathcal{Y}} \frac{\sqrt{n}|(\hat{g}(\theta) - \hat{g}(\theta_y)) - (g(\theta) - g(\theta_y)) - \left(\hat{G}(\theta_y) - G(\theta_y)\right)(\theta - \theta_y)|}{|\theta - \theta_y|} = o_P(1),$$

$$\sup_{|\theta-\theta_y|\leq\delta_n} \sup_{y\in\mathcal{Y}} \frac{\sqrt{n}|\left(\hat{W}(\theta) - \hat{W}(\theta_y)\right) - (W(\theta) - W(\theta_y)) - \left(\hat{H}(\theta_y) - H(\theta_y)\right)(\theta - \theta_y)|}{|\theta - \theta_y|} = o_P(1),$$

and $\sqrt{n}\left(\hat{g}(\theta_y) - g(\theta_y), \hat{G}(\theta_y) - G(\theta_y), \hat{H}(\theta_y) - H(\theta_y)\right) \rightsquigarrow (\mathcal{G}_g(\cdot), \mathcal{G}_G(\cdot), \mathcal{G}_H(\cdot))$.

ASSUMPTION 5 $\sup_{y\in\mathcal{Y}} |y| = o\left(n^{-1/4}\right)$. For any $\delta_n \to 0$,

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

Furthermore, $\hat{W}(\theta) \equiv \hat{W}$, $W(\theta) \equiv W$ and $\hat{W} - W = O_P\left(\frac{1}{\sqrt{n}}\right)$.

Remark 1: We only need one of Assumptions 3, 4 and 5, the last two of which apply to deal with local misspecification that arises in the nonparametric computational approximation in smooth and nonsmooth overidentified models, respectively. Assumption 4 allows for higher order local polynomial regressions but requires $\hat{g}(\cdot)$ and $\hat{W}(\cdot)$ to be multiple times continuously differentiable with a first derivative satisfying a CLT . The restriction this generates is expected from the analysis in Hall and Inoue (2003), and rules out nonparametric style HAC estimates of $\hat{W}(\theta)$. However, Assumption

4 is automatically satisfied in the IL model where the model condition is (asymptotically) linearly separable: $\hat{g}(\theta) = T_n - t(\theta)$, so that (asymptotically) $\hat{G}(\theta) \equiv G(\theta)$ and $\hat{W}(\theta) \equiv W(\theta) = \Sigma(\theta)^{-1}$.

Assumption 5 is used to handle nonsmooth models that involve indicator functions, such as overidentified quantile instrumental variables. See for example, Kim and Pollard (1990). Its current form only allows for local linear regressions and two step style GMM estimates where $\hat{W} = W + O_P\left(\frac{1}{\sqrt{n}}\right)$ (which holds in the quantile IV model of Chernozhukov and Hansen (2005) where the optimal weighting matrix is parameter independent), and rules out nonsmooth continuous updating or other estimates of $\hat{W}$.

Under Assumptions 1 to 5, $\sqrt{n}$ consistency and asymptotic normality of the theoretical posterior mean and distribution, $\bar{\theta}$, $\bar{\eta}$ and the validity of (5) are shown to hold locally uniformly in the addendum, which are important for the local behavior of the feasible estimates $\hat{\theta}$, $\hat{\eta}$ and $\hat{\eta}_\tau$.

ASSUMPTION 6 The kernel function satisfies (1) $\kappa(x) = h(|x|)$ where $h(\cdot)$ decreases monotonically on $(0, \infty)$; (2) $\int \kappa(x)\, dx = 1$; (3) $\int x\kappa(x)\, dx = 0$; (4) $\int |x|^2 \kappa(x)\, dx < \infty$.

THEOREM 1 Under Assumptions 1, 2, 6, and one of 3, 4, or 5, for $\hat{\eta}$ and $\hat{\eta}_\tau$ defined in (6) and (7), or in (14) and (15), both $\sqrt{n}(\hat{\eta} - \bar{\eta}) = o_P^*(1)$ and $\hat{\eta}_\tau - \bar{\eta}_\tau = o_P^*\left(\frac{1}{\sqrt{n}}\right)$ when $Sh^k \to \infty$, $\sqrt{n}h \to \infty$ and $\sqrt{n}h^2 = o(1)$, so that $\hat{\eta}$ and $\hat{\eta}_\tau$ are first order asymptotically equivalent to $\bar{\eta}$ and $\bar{\eta}_\tau$, and posterior inference based on $\hat{\eta}_\tau$ is valid whenever it is valid for the infeasible $\bar{\eta}_\tau$.

In the above we define $X_{n,S} = o_P^*(1)$ if for all $\epsilon, \delta > 0$, $P_n\left(P_{S|n}(|X_{n,S}| \geq \epsilon) > \delta\right) \to 0$ as $n \to \infty$ as in the bootstrap literature, where $P_{S|n}$ is the conditional distribution of the simulation (and S depends on n) and P_n is the distribution of the data. In the appendix, we also denote $X_{n,S} = O_P^*(1)$ if $\forall \delta > 0$, $\exists M > 0$, such that $P_n\left(P_{S|n}(|X_{n,S}| \geq M) > \delta\right) \to 0$.

Two features are worth commenting. First, since the posterior distribution shrinks at $1/\sqrt{n}$ rate, 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 n and h is limited to the "bias" term.

Second, Theorem 1 holds regardless of whether we have exact identification ($d = k$) or overidentification ($d > k$). That the curse of dimensionality is only reflected in k but not d is due to the multicollinearity of moment conditions when $d > k$, in which case the observations Y_n^s are randomly distributed along a manifold of dimension k, and can be handled with a change of variable along this manifold. The lower bound on S is $S >> n^{k/4}$ in the sense that $n^{-k/4}S \to \infty$. The next theorem extends the local linear regression results to more general local polynomial regressions when the moment conditions are either exactly identifying or smooth.

THEOREM 2 Under Assumptions 1, 2 and 6, and one of 3, 4, for $\hat{\eta}$ and $\hat{\eta}_\tau$ defined in (8) and (9), if $nh^{2(p+1)} \to 0$, $\sqrt{n}h \to \infty$, $Sh^k \to \infty$, then $\hat{\theta} - \bar{\theta} = o_P^*(1/\sqrt{n})$, $\hat{\eta} - \bar{\eta} = o_P^*\left(\frac{1}{\sqrt{n}}\right)$, and $\hat{\eta}_\tau - \bar{\eta}_\tau = o_P^*(1/\sqrt{n})$, so that posterior inference based on $\hat{\eta}_\tau$ is valid whenever it is valid for the infeasible $\bar{\eta}_\tau$.

The lower bound on S implied by Theorem 2 is given by $S >> n^{\frac{k}{2(p+1)}}$, which can be much smaller than $S >> n^{k/4}$ by using a larger p. Higher order kernel functions are often used in place of local polynomials for bias reduction in nonparametric regressions. Locally constant (and possibly higher order) kernel mean and quantile estimates of η are as usual given by

$$\hat{\eta} = \sum_{s=1}^{S} \eta^s \kappa\left(\frac{y_n^s}{h}\right) / \sum_{s=1}^{S} \kappa\left(\frac{y_n^s}{h}\right), \tag{20}$$

and

$$\hat{\eta}_\tau = \arg\min_a \sum_{s=1}^{S} \rho_\tau(\eta^s - a)\, \kappa\left(\frac{y_n^s}{h}\right). \tag{21}$$

However, the conditions required for $\sqrt{n}$-consistency and asymptotic normality are substantially more stringent for (20) and (21) as in the following theorem.

THEOREM 3 Under Assumptions 1, 2, 6, and one of 3, 4, or 5, For $\hat{\eta}$ and $\hat{\eta}_\tau$ defined in (20) and (21), $\hat{\eta} - \bar{\eta} = o_P^*\left(\frac{1}{\sqrt{n}}\right)$ and $\hat{\eta}_\tau - \bar{\eta}_\tau = o_P^*\left(\frac{1}{\sqrt{n}}\right)$ if $Sh^k \min\left(1, \frac{1}{nh^2}\right) \to \infty$ and $\sqrt{n}h^2 \to 0$ when $d = k$. The same conclusion holds when $d > k$ under the additional condition that $\sqrt{n}h \to \infty$.

Comparing to Theorem 2, the stricter requirement of $Sh^k \min\left(1, \frac{1}{nh^2}\right) \to \infty$ (than $Sh^k \to \infty$) also implies a larger lower bound on S: $S >> n^{k/4}\sqrt{n}$. This is related to the different bias reduction mechanisms involved between using either a kernel or a local linear term. The (2nd order) kernel reduces bias by leveraging the similarity of the derivatives due to continuity from both sides of the point of interest. The linear regression term instead remove directly the bias associated with the linear term in the Taylor expansion which is suitable for one-sided situations. In conventional local constant kernel regressions where the error variance is $O(1)$, the contribution the bias from the Taylor expansion to the total variance ($O\left(h^2\right)$) is asymptotically negligible. This is no longer the case when the error variance is $O(1/n)$. The total variance is $O\left(\frac{1}{n} + h^2\right)$ when a kernel rather than a linear term is used to reduce bias.

Relatedly, note also that a regular 2nd order kernel is used in Theorem 2. 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

by the same calculation leading to Theorems 2 and 3,

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

Furthermore, $Var\left(\tilde{\theta}-\theta_0\right)=O\left(1/n\right)$. Therefore, the results of Theorem 2 hold under the alternative conditions that $\sqrt{n}h^{\max(p_1+1,p_2)}\to 0$, $Sh^{k}\to\infty$, $\sqrt{n}h\to\infty$, and $\frac{Sh^{k-2(p_1+1)}}{n}\to\infty$. They reduce to Theorem 2 when $p_1+1\geq p_2$. But when $p_2>p_1+1$, it implies that $S>>n^{\frac{k}{2p_2}}n^{\frac{p_2-p_1-1}{p_2}}$, which is strictly stronger than $S>>n^{\frac{k}{2p_2}}$ when a (p_2-1)th order polynomial is used. 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.

5 An illustrative example

We use an example of a simple case of normal sample means to illustrate how the sampling properties of the ABC style nonparametric regressions of θ^s on y_n^s local to zero depend on the conditional distribution of $f\left(\theta^s|y_n^s\right)$ and on the marginal density of $f\left(y_n^s\right)$ for y_n^s close to zero. Unlike conventional nonparametric regressions, both the marginal density of y_n^s and the conditional variance (or conditional density in the quantile regression case) of θ^s given y_n^s are sample size dependent. In particular, define $\bar{\theta}\left(y\right)=E_n\left(\theta^s|y_n^s=y\right)$. It will be shown that under general conditions $f\left(\theta^s|y_n^s=y\right)$ concentrates on a $O\left(\frac{1}{\sqrt{n}}\right)$ neighbourhood of $\bar{\theta}\left(y\right)$. Therefore we expect that $Var\left(\theta^s|y_n^s=y\right)=O\left(1/n\right)$ and that $f\left(\bar{\theta}\left(y\right)|y_n^s=y\right)=O\left(n^{k/2}\right)$. Furthermore, in an exactly identified model $f_{y_n^s}\left(0\right)=O_p\left(1\right)$, while in an overidentified model where $d>k$, $f_{y_n^s}\left(0\right)=O_P\left(n^{\frac{d-k}{2}}\right)$.

To illustrate, consider a vector of sample means in a normal model $\bar{X}_n\sim\left(\mu,\frac{1}{n}\Sigma\right)$, where Σ is known. Let $\theta=\mu$, $\hat{g}\left(\mu\right)=\mu-\bar{X}_n$, and let $\pi\left(\mu\right)=N\left(\mu_0,\Sigma_0\right)$. For $\xi\sim N\left(0,1\right)$, let

$$Y_n^s=\mu^s-\bar{X}_n+\frac{1}{\sqrt{n}}\Sigma^{1/2}\xi.$$

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

$$E\left(\mu^s|Y_n^s=y\right)=\frac{\Sigma}{n}\left(\Sigma_0+\frac{\Sigma}{n}\right)^{-1}\mu_0+\Sigma_0\left(\Sigma_0+\frac{\Sigma}{n}\right)^{-1}\left(\bar{X}_n+y\right)\overset{n\to\infty}{\longrightarrow}\bar{X}_n+y$$

and

$$Var\left(\mu^s|Y_n^s=y\right)=\Sigma_0\left(\Sigma_0+\frac{1}{n}\Sigma\right)^{-1}\frac{\Sigma}{n}=O\left(\frac{1}{n}\right).$$

Under exact identification ($d = k$), whenever Σ_0 is nonsingular, the marginal density of y_n^s is

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

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 $n \to \infty$. Instead, note that

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

In this case,

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

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

$$Var(\mu | Y_n^s = y) = \sigma_0^2 - \sigma_0^4 l' \left(\Sigma_0 + \Sigma/n\right)^{-1} l = \sigma_0^2 \frac{1}{1 + \sigma^2 n l' \Sigma^{-1} l}.$$

The marginal density of Y_n at $t = 0$: $N\left(\bar{X} - u_0 l, \left(\frac{\Sigma}{n} + \sigma_0^2 ll'\right)\right)$ becomes singular when $n \to \infty$. 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 $f_{Y_n^s}(0)$ under correct specification becomes

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

It is also easy to show that

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

using the relation that $\det(I + uv') = 1 + u'v$. If the model is incorrectly specified, $\Delta \to \infty$, $f_{Y_n^s}(0)$ becomes exponentially small. The general result mirrors this example.

LEMMA 1 Let $\theta = \arg\min g\left(\tilde{\theta}\right)' \hat{W}\left(\tilde{\theta}\right) \hat{g}\left(\tilde{\theta}\right)$. Under Assumptions 1 and 2, and one of 3, 4 and 5, for $W = W(\theta_0)$, $f_{Y_n}(0)/\sqrt{n}^{d-k} f_\infty(0) \xrightarrow{p} 1$, where

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

This lemma includes several possibilities. In an exactly identified model where $d = k$, and $\hat{g}\left(\tilde{\theta}\right) = 0$, $f_{Y_n^s}(0) = O_P(1)$. In a correctly specified and overidentified model, $\hat{g}\left(\tilde{\theta}\right) = O_P(1/\sqrt{n})$, $f_{Y_n^s}(0) = O_P\left(\sqrt{n}^{d-k}\right)$. If the overidentified model is misspecified, $f_{Y_n^s}(0)$ is exponentially small when the sample size n increases: $f_{Y_n^s}(0) = O_p(\exp(-nc))$ for some $c > 0$.

Next to illustrate that the singularity of $f(Y)$ can be handled through a change of variable, consider for simplicity $\Sigma = I$. Partition $Y = (Y_1, Y_2)$ for a scalar Y_1. Let $Y_2 = \ell Y_1 + \frac{\Delta}{\sqrt{n}} + \frac{w_2}{\sqrt{n}}$. Then $Y_1 = \mu - \bar{X}_1 + \frac{\xi_1}{\sqrt{n}}$, $Y_2 = \mu - \bar{X}_2 + \frac{\xi_2}{\sqrt{n}}$, $\Delta = -\sqrt{n}\left(\bar{X}_2 - \bar{X}_1\right) = O_p(1)$, and $w_2 = \xi_2 - \xi_1 = O_p(1)$. The implication of this for the kernel function is that

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

If $\sqrt{n}h \to \infty$, then $\frac{\Delta}{\sqrt{n}h} = o_p(1)$, $\frac{w_2}{\sqrt{n}h} = o_P(1)$, 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 the more general nonlinear setting which is used in the proof in the appendix.

With this change of variable in the normal example we write $\mu = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \epsilon$, 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.

6 Conclusion

We build on previous works by Creel and Kristensen (2011), Chernozhukov and Hong (2003) and Gao and Hong (2014) and provide a careful asymptotic analysis of the Bayesian Indirect Inference method. We show that local linear and polynomial estimators have theoretical advantages over kernel methods, and that Generalized Method of Moment models can also be computed by ABC style methods. In future work we plan to investigate sieve implementation of BIL and ABC-GMM models, and to validate bootstrap and other resampling methods in this context. Local polynomial methods are known to achieve optimal rates in estimating nonparametric functions. It remains to be seen whether this holds for the BIL and ABC-GMM models and whether sieve methods can be a viable contender.

References

An, S. and F. Schorfheide (2007): "Bayesian Analysis of DSGE Models," *Econometric Reviews*, 26, 113–172.

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

Beaumont, M. A., W. Zhang, and D. J. Balding (2002): "Approximate Bayesian computation in population genetics," *Genetics*, 162, 2025–2035.

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

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

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

Creel, M. and D. Kristensen (2015): "On selection of statistics for approximate Bayesian computing (or the Method of Simulated Moments)," *Computational Statistics & Data Analysis*.

Creel, M. D. and D. Kristensen (2011): "Indirect likelihood inference," Working paper, available at http://ddd.uab.cat/record/71449/.

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

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2005): "Estimating dynamic equilibrium economies: linear versus nonlinear likelihood," *Journal of Applied Econometrics*, 20, 891–910.

Forneron, J.-J. and S. Ng (2015): "The ABC of Simulation Estimation with Auxiliary Statistics," *arXiv preprint arXiv:1501.01265*.

Gallant, A. R. and H. Hong (2007): "A statistical inquiry into the plausibility of recursive utility," *Journal of Financial Econometrics*, 5, 523–559.

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

Gao, J. and H. Hong (2014): "A Computational implementation of GMM," Available at SSRN working paper 2503199, http://ssrn.com/abstract=2503199.

Gentzkow, M. and J. M. Shapiro (2014): "Measuring the sensitivity of parameter estimates to sample statistics," Tech. rep., National Bureau of Economic Research, nBER, available at http://www.nber.org/papers/w20673/.

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

Guerrón-Quintana, P. A. (2010): "What you match does matter: the effects of data on DSGE estimation," *Journal of Applied Econometrics*, 25, 774–804.

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, 361–394.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

Jun, S. J., J. Pinkse, and Y. Wan (2011): "-Consistent robust integration-based estimation," *Journal of Multivariate Analysis*, 102, 828–846.

——— (2015): "Classical Laplace estimation for-consistent estimators: Improved convergence rates and rate-adaptive inference," *Journal of Econometrics*, 187, 201–216.

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

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

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, North Holland, 2113–2241.

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

Pollard, D. (1991): "Asymptotics for Least Absolute Deviation Regression Estimator," *Econometric Theory*, 7, 186–199.

Press, W., S. A. Teukolsky, W. Vettering, and B. Flannery (1992): *Numerical Recipes in C, The Art of Scientific Computing*, Cambridge.

Robinson, P. M. (1988): "The stochastic difference between econometric statistics," *Econometrica: Journal of the Econometric Society*, 531–548.

Ruge-Murcia, F. (2012): "Estimating nonlinear DSGE models by the simulated method of moments: With an application to business cycles," *Journal of Economic Dynamics and Control*, 36, 914–938.

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

Table 1: DSGE models, support of uniform priors.

Parameter	Lower bound	Upper bound	True value	Prior bias	Prior RMSE
α	0.2	0.4	0.330	-0.030	0.065
β	0.95	1	0.990	-0.015	0.021
δ	0.01	0.1	0.025	0.030	0.040
γ	0	5	2.000	0.500	1.527
ρ_z	0	1	0.900	-0.400	0.493
σ_z	0	0.1	0.010	0.030	0.042
ρ_η	0	1	0.700	-0.200	0.351
σ_η	0	0.1	0.005	0.040	0.049
$\bar{n}$	6/24	9/24	1/3	-0.021	0.042

Table 2: Selected statistics, DSGE model. For statistics 11-20, σ_{xy} indicates the sample covariance of the residuals of the AR1 models for the respective variables x and y.

Statistic	Description	Statistic	Description
1	$\widehat{\log \psi}$ from eq. 17	12	σ_{qq}
2	$\widehat{\gamma}$ from eq. 17	13	σ_{qn}
3	$\widehat{\rho}_{\eta}$, residuals of eq. 17	14	σ_{qr}
4	sample mean c	15	σ_{qw}
5	sample mean n	16	σ_{cc}
6	sample std. dev. q	17	σ_{cn}
7	sample std. dev. c	18	σ_{cr}
8	sample std. dev. n	19	σ_{cw}
9	sample std. dev. r	20	σ_{nn}
10	sample std. dev. w	21	σ_{ww}
11	estimated AR1 coef., r	22	$\overline{c/n}$

Table 3: DSGE model. Monte Carlo results (1000 replications). Bandwidths tuned using prior. LC=local constant, LL=local linear, LQ=local quadratic. 90% CI gives the proportion of times that the true value is in the 90% confidence interval.

	Bias			RMSE			90% CI
Parameter	LC	LL	LQ	LC	LL	LQ	LC
α	0.025	0.002	0.001	0.032	0.013	0.012	0.920
β	-0.008	0.001	0.001	0.010	0.003	0.003	0.993
δ	0.007	0.001	-0.000	0.011	0.004	0.003	0.991
γ	0.037	0.037	0.006	0.158	0.103	0.106	0.986
ρ_z	-0.012	-0.003	0.001	0.040	0.012	0.009	0.877
σ_z	-0.001	-0.001	-0.000	0.003	0.002	0.002	0.893
ρ_η	-0.007	-0.011	-0.009	0.054	0.047	0.049	1.000
σ_η	0.001	-0.000	0.000	0.003	0.002	0.001	0.834
$\bar{n}$	0.003	0.001	0.001	0.005	0.004	0.004	0.731

Table 4: DSGE model. Monte Carlo results (1000 replications). Bandwidths tuned locally. LC=local constant, LL=local linear, LQ=local quadratic. 90% CI gives the proportion of times that the true value is in the 90% confidence interval.

	Bias			RMSE			90% CI
Parameter	LC	LL	LQ	LC	LL	LQ	LC
α	0.027	0.003	0.001	0.033	0.013	0.012	0.916
β	-0.008	0.001	0.002	0.011	0.003	0.003	1.000
δ	0.008	0.001	-0.000	0.011	0.004	0.003	0.900
γ	0.031	0.036	0.005	0.145	0.103	0.099	0.922
ρ_z	-0.013	-0.002	0.001	0.040	0.010	0.008	0.900
σ_z	-0.001	-0.001	-0.008	0.003	0.002	0.002	0.863
ρ_η	-0.010	-0.012	-0.010	0.054	0.046	0.049	0.794
σ_η	0.001	0.000	0.000	0.003	0.002	0.001	0.835
$\bar{n}$	-0.006	0.001	0.002	0.006	0.004	0.004	0.921

Table 5: Quantile IV model. Monte Carlo results (1000 replications). Bandwidths tuned using prior. LC=local constant, LL=local linear. 90% CI gives the proportion of times that the true value is in the 90% confidence interval.

		β_1	β_2
Bias	Prior	0.5	0.5
	IV	0.104	0.229
	LC	0.005	0.008
	LL	0.003	0.006
RMSE	Prior	1.0	1.0
	IV	0.107	0.232
	LC	0.023	0.045
	LL	0.019	0.038
90% CI	LC	0.858	0.903

Table 6: Quantile IV model. Monte Carlo results (1000 replications). Bandwidths tuned locally. LC=local constant, LL=local linear. 90% CI gives the proportion of times that the true value is in the 90% confidence interval.

		β_1	β_2
Bias	LC	0.009	0.018
	LL	0.005	0.010
RMSE	LC	0.028	0.056
	LL	0.019	0.038
90% CI	LC	0.899	0.912

A Proofs of Theorems

A.1 Proof of Theorem 1

Consider first the *local linear mean* $\hat{\eta}$ in (6) and (14). Define $\kappa_s = \kappa\left(\frac{y_n^s}{h}\right)$ and $Z_s = \left(1, \frac{y_n^s}{h}\right)'$. Furthermore, let $m(y) = \bar{\eta}^y = E(\eta|Y_n^s = y)$ (see Lemma 5), $a_0 = m(0)$ and $b_0 = m'(0)$, which is defined by finite differencing (68) along a sequence $\epsilon_n \to 0$ and $\sqrt{n}\epsilon_n \to \infty$ so that $b_0 = \frac{\partial}{\partial y}\eta(\theta_y)|_{y=0} + o_P(1)$. Or simply let $b_0 = \eta'(0)$. Also let $\eta_s^* = \eta^s - a_0 - b_0 y_n^s$. Then one can write

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

We separately consider the exact identification case and the overidentification case.

Exact Identification Consider first H. By Lemma 1, $f_{Y_n}(y) = f_{Y_n}^{\infty}(y)(1 + o_P(1))$. Let E_n denote the conditional distribution given the data (with respect to $\pi(\theta)$ and the residual variance ξ in (13)), and Var_n the corresponding conditional variance given the data. Then

$$\begin{aligned} E_n H &= \int (1\ v)(1\ v)' \kappa(v) f_{Y_n}(vh)\, dv \\ &= (1 + o_P(1)) \int (1\ v)(1\ v)' \kappa(v) f_{Y_n}^{\infty}(vh)\, dv = (1 + o_P(1)) f_{Y_n}^{\infty}(0) C_\kappa + o_P(1), \end{aligned}$$

where $C_\kappa = \int (1\ v)(1\ v)' \kappa(v)\, dv$. Next, for $i, j = 0, \ldots, k$, with $v_0 = v_1 = 1$,

$$\begin{aligned} &Var_n\left(\sqrt{Sh^k} H_{ij}\right) = \frac{1}{h^k} Var_n\left(Z_{s,i} Z_{s,j} \kappa_s\right) = E_n \frac{1}{h^k} Z_{s,i}^2 Z_{s,j}^2 \kappa_s^2 - h^k (E_n H)^2 \\ &= \int v_i^2 v_j^2 \kappa^2(v) f_{Y_n}(vh)\, dv - h^k (E_n H)^2 = (1 + o_P(1)) f_{Y_n}^{\infty}(0) \int v_i^2 v_j^2 \kappa^2(v)\, dv + o_P(1) - h^k (E_n H)^2, \end{aligned}$$

so that $Var_n(H_{ij}) = o_P(1)$ and hence $H = f_{Y_n}^{\infty}(0) C_\kappa + o_P^*(1)$. Next consider the bias and variance of J separately. Consider the bias first. Using (68) we can write

$$\begin{aligned} E_n J &= \frac{\sqrt{n}}{h^k} E_n Z_s \kappa_s \left(E(\eta|y^s) - a_0 - b_0 y^s\right) = \frac{\sqrt{n}}{h^k} E_n Z_s \kappa_s \left(\frac{1}{2} y_s' \eta''(0) y_s + O\left(y_s^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\ &= \sqrt{n} \int (1\ v)' \kappa(v) \left(h^2 \frac{1}{2} v' \eta''(0) v + O\left(v^3 h^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) f_{Y_n}(vh)\, dv \\ &= (1 + o_P(1)) f_{Y_n}^{\infty}(0) \left[\sqrt{n} h^2 \int (1\ v)' \frac{v' \eta''(0) v}{2} \kappa(v)\, dv + o_P(1) \int (1\ v)' \kappa(v)\, dv\right] + o_P(1) = o_P(1). \end{aligned}$$

Next consider the following decomposition of the variance,

$$Var_n\left(\frac{\sqrt{n}}{\sqrt{Sh^k}} \sum_{s=1}^{S} Z_s \kappa_s \eta_s^*\right) = \frac{n}{h^k} Var_n\left(Z_s \kappa_s \eta_s^*\right) = \frac{n}{h^k}\left[E_n Var_n\left(Z_s \kappa_s \eta_s^* | y_s\right) + Var_n E_n\left(Z_s \kappa_s \eta_s^* | y_s\right)\right]. \tag{22}$$

For the first term, by (65), $Var_n\left(\sqrt{n}\eta|y^s\right) = J\left(y^s\right)^{-1} + o_P\left(1\right)$ uniformly in $y \in \mathcal{Y}$,

$$\begin{aligned}
&\frac{n}{h^k}E_nVar_n\left(Z_s\kappa_s\eta_s^*|y_s\right) = \frac{1}{h^k}E_nZ_sZ_s'\kappa_s^2Var_n\left(\sqrt{n}\eta^s|y^s\right) \\
=&\frac{1}{h^k}\int Z_sZ_s'\kappa_s^2\left[J\left(y_s\right)^{-1} + o_P\left(1\right)\right]f_{Y_n}\left(y_s\right)dy_s = J\left(0\right)^{-1}f_{Y_n}^{\infty}\left(0\right)\int\left(1\ v\right)'\left(1\ v\right)\kappa^2\left(v\right)dv + o_P\left(1\right).
\end{aligned}$$

For the second term,

$$\begin{aligned}
&\frac{n}{h^k}Var_nZ_s\kappa_sE_n\left(\eta_s^*|y_s\right) = \frac{n}{h^k}Var_nZ_s\kappa_s\left(\frac{1}{2}y_s'\eta''\left(0\right)y_s + O\left(y_s^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \frac{1}{h^k}E_nZ_sZ_s'\kappa_s^2\left(\sqrt{n}\frac{1}{2}y_s'\eta''\left(0\right)y_s + \sqrt{n}O\left(y_s^3\right) + o_P\left(1\right)\right)^2 - h^k\left(E_n\frac{\sqrt{n}}{h^k}Z_s\kappa_s\eta_s^*\right)^2 \\
&= nh^4f_{Y_n}^{\infty}\left(0\right)\int\left(1\ v\right)'\left(1\ v\right)\kappa^2\left(v\right)\left(\frac{1}{2}v'\eta''\left(0\right)v\right)^2dv + o_P\left(1\right) + h^k\left(E_nJ\right)^2 = o_P\left(1\right).
\end{aligned}$$

Therefore Since $Sh^k \to \infty$, these calculations show that $Var_nJ = o_P\left(1\right)$. Therefore $J = o_P^*\left(1\right)$. Essentially, we have shown that $J = O_P^*\left(\frac{1}{\sqrt{Sh^k}}\left(1+\sqrt{n}h^2\right) + \sqrt{n}h^2\right)$. By the definition of $f_{Y_n}^{\infty}\left(y\right)$ in Lemma 1, $f_{Y_n}^{\infty}\left(y\right)^{-1} = O_P\left(1\right)$ since $-n\hat{Q}_y\left(\theta_y\right) = O_P\left(1\right)$. Then we can write $H^{-1}J = \left(C_\kappa^{-1}\left(f_{Y_n}^{\infty}\left(0\right)\right)^{-1} + o_P^*\left(1\right)\right)J = o_P^*\left(1\right)$.

Over Identification In this case the asymptotic distribution of the regressors Y_s are collinearly centered along a $d-k$ dimensional manifold with variance of the order $O\left(1/n\right)$. 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.

To begin with, partition $Y = \left(Y_1, Y_2\right)$ where $Y_1 \in R^k, Y_2 \in R^{d-k}$, and the population moments $g\left(\theta\right)$ (where $g\left(\theta\right) = t\left(\theta_0\right) - t\left(\theta\right)$ in BIL) correspondingly into $g_1\left(\theta\right), g_2\left(\theta\right)$. Define $\Delta_n = y_2 - g_2\left(g_1^{-1}\left(y_1\right)\right)$, so that $\Delta_n = O_P\left(1\right)$ since

$$\begin{aligned}
\Delta_n &= \sqrt{n}\left(\hat{g}_2\left(\theta\right) + \frac{\epsilon_2}{\sqrt{n}} - \left(g_2\left(g_1^{-1}\left(\hat{g}_1\left(\theta\right) + \frac{\epsilon_1}{\sqrt{n}}\right)\right)\right)\right) \\
&= \sqrt{n}\left(\hat{g}_2\left(\theta\right) - g_2\left(\theta\right)\right) + \epsilon_2 - G_2\left(\theta\right)G_1\left(\theta\right)^{-1}\sqrt{n}\left(\hat{g}_1\left(\theta\right) - g_1\left(\theta\right) + \frac{\epsilon_1}{\sqrt{n}}\right) + o_P\left(1\right).
\end{aligned}$$

Also define $c = G_2\left(\theta_0\right)G_1\left(\theta_0\right)^{-1}$, so that $\lim_{h\to} g_2\left(g_1^{-1}\left(uh\right)\right)/h = cu$. Consider the change of variable $\left(y_1, y_2\right) \to \left(w_1, w_2\right)$, where

$$w_1 = y_1 \quad w_2 = \sqrt{n}\left(y_2 - cy_1\right) \quad \text{so that} \quad w_2 = \sqrt{n}\left(g_2\left(g_1^{-1}\left(y_1\right)\right) - cy_1 + \frac{\Delta_n}{\sqrt{n}}\right). \tag{23}$$

Then we can define $f_W\left(w\right) = \sqrt{n}^{d-k}f_{Y_n}\left(w_1, cw_1 + \frac{w_2}{\sqrt{n}}\right)$. Rewrite the regression function as $\eta = a + b'y = a + w_1'd_1 + w_2'd_2 = a + w'd$, for $d_1 = \left(b_1 + c'b_2\right)$ and $d_2 = b_2/\sqrt{n}$. Define as before

$\kappa_s = \kappa\left(\frac{y^s}{h}\right)$ and $Z_s = \left(1, \frac{w_1^s}{h}, w_2^s\right)'$. Furthermore, $m(y) = E(\eta|Y=y)$, $a_0 = m(0)$, $b_0 = \eta'(0)$. Also let $\eta_s^* = \eta^s - a_0 - b_0 y^s = \eta^s - a_0 - d_0 w^s$. Then, write

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

Also define $f_W^\infty(w) = f_{Y_n}^\infty\left(w_1, cw_1 + \frac{w_2}{\sqrt{n}}\right)$. Note that we can also replace $\hat{W}(\theta)$ by $W = W(\theta_0)$ in $n\hat{Q}_y\left(\check{\theta}_y\right)$ in (55) (and absorbed into $(1 + o_P(1))$). For $\bar{C}_y$ in (55), write

$$f_W^\infty(w) = \bar{C}_{y(w)} e^{-n\left(y(w) - \hat{g}\left(\check{\theta}_{y(w)}\right)\right)' W\left(y(w) - \hat{g}\left(\check{\theta}_{y(w)}\right)\right)} \quad \text{where} \quad y(w) = \left(w_1, cw_1 + w_2/\sqrt{n}\right), \tag{24}$$

Note that uniformly in w, $\sqrt{n}\left(w_1 - \hat{g}_1\left(\check{\theta}_{y(w)}\right)\right) = O_p(1)$, and

$$\begin{aligned} cw_1 - \hat{g}_2\left(\check{\theta}_{y(w)}\right) =& cw_1 - g_2\left(g_1^{-1}(w_1)\right) + g_2\left(g_1^{-1}(w_1)\right) - g_2\left(\theta_{y(w)}\right) + O_P\left(\frac{1}{\sqrt{n}}\right) \\ =& -\frac{\partial g_2}{\partial \theta}\frac{\partial \theta}{\partial w_2}\frac{w_2}{\sqrt{n}} + O\left(\left(\frac{w_2}{\sqrt{n}}\right)^2\right) + O\left(w_1^2\right) + O_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

This can be used to show that for a positive definite and definite matrix C_{22},

$$-n\left(y(w) - \hat{g}\left(\check{\theta}_{y(w)}\right)\right)' W\left(y(w) - \hat{g}\left(\check{\theta}_{y(w)}\right)\right) = -\left(w_2 - O_P(1)\right)' C_{22}\left(w_2 - O_P(1)\right) + O_P(1). \tag{25}$$

Consider then first H. Note that $\kappa\left(\frac{y}{h}\right) = \kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right)$. Write, using $\sqrt{n}h \to \infty$,

$$\begin{aligned} E_n H =& \int \frac{Z_s Z_s' \kappa_s f_W(w)}{h^k} dw = \frac{1 + o_P(1)}{h^k}\int \begin{pmatrix} 1 \\ \frac{w_1}{h} \\ w_2 \end{pmatrix}\left(1 \ \frac{w_1}{h} \ w_2\right)\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) f_W^\infty(w)\, dw \\ =& (1 + o_P(1))\int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix}(1 \ u \ w_2)\,\kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right)\bar{C}_{y(uh, w_2)} e^{-(w_2 - O_P(1))' C_{22}(w_2 - O_P(1)) + O_P(1)} du dw_2 \\ =& (1 + o_P(1))\,\bar{C}_0 \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix}(1 \ u \ w_2)\,\kappa(u, cu)\, e^{-(w_2 - O_P(1))' C_{22}(w_2 - O_P(1)) + O_P(1)} du dw_2 + o_P(1) \\ =& (1 + o_P(1))\,\hat{H}_n + o_P(1) \quad \text{for} \quad \hat{H}_n \quad \text{positive definite w.p} \to 1. \end{aligned} \tag{26}$$

Similar calculations can also be used to check that $Var_n(H) = o_P(1)$. Therefore $H = \hat{H}_n + o_P^*(1)$.

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

$$E_n J = \frac{\sqrt{n}}{h^k} E Z_s \kappa_s \left(E\left(\eta | y^s\right) - a_0 - b_0 y^s\right) = \frac{\sqrt{n}}{h^k} E_n Z_s \kappa_s \left(\frac{1}{2} y_s' \eta''(0) y_s + O\left(y_s^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right)$$

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

$$\left[\frac{1}{2} \begin{pmatrix} uh \\ cuh + \frac{w_2}{\sqrt{n}} \end{pmatrix}' \eta''(0) \begin{pmatrix} uh \\ cuh + \frac{w_2}{\sqrt{n}} \end{pmatrix} + o\left(u^2 h^2 + \frac{w^2}{n}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right] f_W(uh, w_2)\, du dw_2$$

$$= \sqrt{n} h^2 \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) \frac{1}{2} \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix}' \left(\eta''(0) + o_P(1)\right) \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix} f_W^{\infty}(0, w_2)\, du dw_2 + o_P(1)$$

$$= \sqrt{n} h^2 \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \kappa(u, cu) \frac{1}{2} \begin{pmatrix} u \\ cu \end{pmatrix}' \left(\eta''(0) + o_P(1)\right) \begin{pmatrix} u \\ cu \end{pmatrix} f_W^{\infty}(0, w_2)\, du dw_2 \left(1 + o_P(1)\right) + o_P(1)$$

Using the form of $f_W^{\infty}(w)$ in (24) and (25), we can declare that $E_n J = O_P\left(\sqrt{n}h^2\right) = o_P(1)$.

The variance also has two terms, as in (22). The first term in variance,

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

$$= \frac{n}{h^k} \int \begin{pmatrix} 1 \\ \frac{w_1}{h} \\ w_2 \end{pmatrix} \begin{pmatrix} 1 & \frac{w_1}{h} & w_2 \end{pmatrix} \kappa^2\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) \times Var\left(\eta_s^* | y_1 = w_1, y_2 = cw_1 + \frac{w_2}{\sqrt{n}}\right) f_W(w)\, dw$$

$$= \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \begin{pmatrix} 1 & u & w_2 \end{pmatrix} \kappa^2\left(u, cu + \frac{w_2}{\sqrt{n}h}\right)\ n\ Var_n\left(\eta_s^* | y_1 = uh, y_2 = cuh + \frac{w_2}{\sqrt{n}}\right) f_W(uh, w_2)\, du dw_2$$

$$= \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \begin{pmatrix} 1 & u & w_2 \end{pmatrix} \left(\kappa^2(u, cu) + o_{a.s}(1)\right) \left(\mathcal{J}^{-1} + o_P(1)\right) f_W^{\infty}(uh, w_2) \left(1 + o_P(1)\right) du dw_2$$

$$= \mathcal{J}^{-1} \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \begin{pmatrix} 1 & u & w_2 \end{pmatrix} \kappa^2(u, cu) f_W^{\infty}(0, w_2)\, du dw_2 + o_P(1) = O_P(1).$$

The second term in variance,

$$\frac{n}{h^k}Var_n Z_s\kappa_s E\left(\eta_s^*|y_s\right)=\frac{n}{h^k}Var_n Z_s\kappa_s\left(\frac{1}{2}y_s'\eta''\left(0\right)y_s+O\left(y_s^3\right)+o_P\left(\frac{1}{\sqrt{n}}\right)\right)$$

$$\leq\frac{1}{h^k}E_n Z_s Z_s'\kappa_s^2\left(\sqrt{n}\frac{1}{2}y_s'\eta''\left(0\right)y_s+\sqrt{n}O\left(y_s^3\right)+o_P\left(1\right)\right)^2$$

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

$$\left(\frac{1}{2}\begin{pmatrix}n^{1/4}w_1\\ cn^{1/4}w_1+\frac{w_2}{n^{1/4}}\end{pmatrix}'\eta''\left(0\right)\begin{pmatrix}n^{1/4}w_1\\ cn^{1/4}w_1+\frac{w_2}{n^{1/4}}\end{pmatrix}+\sqrt{n}O\left(w_1^3+\frac{w_2^3}{n\sqrt{n}}\right)+o_P\left(1\right)\right)^2 f_W\left(w_1,w_2\right)dw_1dw_2$$

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

$$\left(\frac{1}{2}\begin{pmatrix}u\\ cu+\frac{w_2}{\sqrt{n}h}\end{pmatrix}'\eta''\left(0\right)\begin{pmatrix}u\\ cu+\frac{w_2}{\sqrt{n}h}\end{pmatrix}+o\left(u^2+\frac{w_2^2}{nh^2}\right)+o_P\left(1\right)\right)^2 f_W\left(uh,w_2\right)dudw_2$$

$$=nh^4\left[\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\begin{pmatrix}1 & u & w_2\end{pmatrix}\kappa^2\left(u,cu\right)\left(\frac{1}{2}\begin{pmatrix}u\\ cu\end{pmatrix}'\eta''\left(0\right)\begin{pmatrix}u\\ cu\end{pmatrix}\right)^2\left(1+o_P\left(1\right)\right)f_W\left(0,w_2\right)dudw_2+o_P\left(1\right)\right]$$

Then since $nh^4\to 0$ and $Sh^k\to\infty$, we conclude that $Var_nJ=\frac{1}{Sh^k}\frac{n}{h^k}Var_n\left(Z_s\kappa_s\eta_s^*\right)=o_P\left(1\right)$, so that $J=o_P^*\left(1\right)$ and $H^{-1}J=\left(\left(1+o_P\left(1\right)\right)\hat{H}_n+O_P^*\left(1\right)\right)^{-1}o_P^*\left(1\right)=o_P^*\left(1\right)$. In other words,

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

The rate normalization for b_2 depends on the variation of $g_2\left(g_1^{-1}\left(y_1\right)\right)-cy_1=O\left(h^2\right)$, $\hat{g}_2\left(\theta\right)-g_2\left(\theta\right)=O_P\left(\frac{1}{\sqrt{n}}\right)$, $\hat{g}_1\left(\theta\right)-g_1\left(\theta\right)=O_P\left(\frac{1}{\sqrt{n}}\right)$, where the later terms prevails. If $\sqrt{n}h=O\left(1\right)$ instead of $\sqrt{n}h^2\to\infty$, b_2 needs to be normalized by h instead of $1/\sqrt{n}$, and the convergence rate slows to h^{d+k} from h^k.

Local linear quantile regression Consider $\hat{\eta}_\tau=\hat{a}$ defined in (7). We adapt and revise the local linear robust regression method of Fan et al. (1994) to our settings. Extensions to local polynomials are immediate. Recall that $\eta^s=\eta\left(\theta^s\right)$ for a known $\eta\left(\cdot\right):R^k\to R$. The goal is to conduct inference on $\eta_0=\eta\left(\theta_0\right)$. We also discuss the exact and over identification cases separately.

Exact Identification: $d = k$. Let $a_0 = \bar{\eta}_\tau(y=0)$, $b_0 = \frac{\partial}{\partial y}\eta(0)$ (see eqs (67) and (69)), and $Z_s = \left(1, \frac{y_s}{h}\right)$. Define $\hat{\theta} = \sqrt{n}\sqrt{Sh^k}\left(\hat{a} - a_0, h\left(\hat{b} - b_0\right)\right)$. Let $\eta_s^* = \eta_s - a_0 - b_0' y_s$, $\kappa_s = \kappa\left(\frac{y_s}{h}\right)$. Then,

$$\hat{\theta} = \arg\min G_S(\theta) = \sqrt{n}\sum_{s=1}^{S}\left(\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau(\eta_s^*)\right)\kappa_s. \tag{27}$$

Consider now the following decomposition, for $Y = (y_1, \ldots, y_s)$, and for $\rho_\tau'(\cdot) = \tau - 1(\cdot \leq 0)$,

$$G_S(\theta) = E_n(G_S(\theta)|Y) + \left(Sh^k\right)^{-1/2}\sum_{s=1}^{S}\left(\rho_\tau'(\eta_s^*) Z_s'\kappa_s - E_n\left(\rho_\tau'(\eta_s^*)|y_s\right) Z_s'\kappa_s\right)\theta + R_S(\theta), \tag{28}$$

We focus on kernel functions with bounded support (by M). First, by eq (69),

$$\begin{aligned} Q_\tau(\eta_s^*|y_s) =& Q_\tau(\eta_s|y_s) - a_0 - b_0 y_s = \bar{\eta}_\tau^{y_s} - \bar{\eta}_\tau^{y=0} - b_0' y_s \\ =& \eta_y - \eta_0 - b_0' y_s + o_P\left(\frac{1}{\sqrt{n}}\right) = O\left(|y_s|^2\right) + o_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Now write, for $\hat{H}_n \equiv \frac{1}{\sqrt{n}Sh^k}\sum_{s=1}^{S} E_n\left(\rho_\tau''(\eta_s^*)|y_s\right) Z_s Z_s'\kappa_s$,

$$\begin{aligned} E_n(G_S(\theta)|Y) =& \sqrt{n}\sum_{s=1}^{S} E_n\left[\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau(\eta_s^*)\middle| y_s\right]\kappa_s \\ =& \left(Sh^k\right)^{-1/2}\sum_{s=1}^{S} E_n\left(\rho_\tau'(\eta_s^*)|y_s\right) Z_s'\kappa_s\theta + (1 + o_P(1))\frac{1}{2}\theta'\hat{H}_n\theta. \end{aligned} \tag{29}$$

With $f_{\eta_s^*}^n(\cdot|y_s)$ the conditional density of η_s^* given y_s, *conditional on the data*,

$$E_n\left(\rho_\tau''(\eta_s^*)|y_s\right) = f_{\eta_s^*}^n(0|y_s) = f_{\eta_s}^n(a_0 + b_0 y_s|y_s) = \sqrt{n} f_{\sqrt{n}(\eta_s - \eta(\check{\theta}_{y_s}))}^n\left(\sqrt{n}\left(a_0 + b_0 y_s - \eta\left(\check{\theta}_{y_s}\right)\right)|y_s\right)$$

By finite differencing (73) (wrt to s) along a sequence converging to 0 sufficiently slowly,

$$\sup_s\left|f_{\sqrt{n}(\eta_s - \eta(\check{\theta}_{y_s}))}^n(s) - \phi\left(s; \Sigma_{\eta(y)}\right)\right| = o_P(1).$$

where $\Sigma_{\eta(y)} \equiv \frac{\partial}{\partial\theta}\eta(\theta_y)' J_y^{-1}\frac{\partial}{\partial\theta}\eta(\theta_y)$, and $\phi\left(s; \Sigma_{\eta(y)}\right) = \frac{1}{\sqrt{\Sigma_{\eta(y)}}}\phi\left(\frac{s}{\sqrt{\Sigma_{\eta(y)}}}\right)$. Furthermore,

$$\begin{aligned} \sqrt{n}\left(a_0 + b_0 y_s - \eta\left(\check{\theta}_{y_s}\right)\right) =& \sqrt{n}\left(\bar{\eta}_\tau^0 - \eta(\theta_0)\right) + \sqrt{n}\left(\eta(\theta_0) + b_0 y_s - \eta(\theta_{y_s})\right) + \sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right)\right) \\ =& q_\tau\sqrt{\Sigma_{\eta(0)}} + o_P(1) + \sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right)\right) \end{aligned}$$

using (67), $a_0 + b_0 y_s - \eta(\theta_{y_s}) = O\left(h^2\right) + o_P\left(\frac{1}{\sqrt{n}}\right)$ and $\sqrt{n}h^2 = o(1)$. Therefore we can write

$$\frac{E_n\left(\rho_\tau''(\eta_s^*)|y_s\right)}{\sqrt{n}} = \phi\left(q_\tau\sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right)\right) + o_P(1); \Sigma_{\eta(y)}\right) + o_P(1) \tag{30}$$

Therefore by the usual change of variable $y_s = vh$, for $C_\kappa = \int (1\ v)' (1\ v) \kappa(v)\, dv$, and recalling $\sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right) - \eta(\theta_0) - \eta\left(\check{\theta}_0\right)\right) = o_P(1)$ as $y_s \to 0$,

$$\begin{aligned} E_n \hat{H}_n \equiv & E_n \frac{1}{Sh^k} \sum_{s=1}^{S} \frac{E_n\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)}{\sqrt{n}} Z_s Z_s' \kappa_s = \int \frac{1}{h^k} \frac{E_n\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)}{\sqrt{n}} Z_s Z_s' \kappa_s f_{Y_n}(y_s)\, dy_s \\ = & \int \frac{E_n\left(\rho_\tau''\left(\eta_s^*\right)|y_s = vh\right)}{\sqrt{n}} (1\ v)' (1\ v) \kappa(v) f_{Y_n}(vh)\, dv \\ = & f_{Y_n}^{\infty}(0)\, \phi\left(q_\tau \sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta(\theta_0) - \eta\left(\check{\theta}_0\right)\right); \Sigma_{\eta(0)}\right) C_\kappa + o_P(1). \end{aligned}$$

The same calculation also shows that

$$Var_n \frac{1}{Sh^k} \sum_{s=1}^{S} \frac{E_n\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)}{\sqrt{n}} Z_s Z_s' \kappa_s = o_P(1). \tag{31}$$

Therefore, $\hat{H}_n = H + o_P^*(1)$, where $H = f_{Y_n}^{\infty}(0)\, \phi\left(q_\tau \sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta(\theta_0) - \eta\left(\check{\theta}_0\right)\right); \Sigma_{\eta(0)}\right) C_\kappa$. Hence we can write, for θ such that $|\theta| / \left(\sqrt{n}\sqrt{Sh^k}\right) \to 0$,

$$G_S(\theta) = \frac{1}{2} \theta' \left(H + o_P^*(1)\right) \theta + \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} \rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s \theta + R_S(\theta). \tag{32}$$

Next we show that $R_S(\theta) = o_P^*(1)$ for fixed θ. Since $E_n R_S(\theta) = 0$, it suffices to bound $E_n R_S(\theta)^2$,

$$\begin{aligned} E_n R_S(\theta)^2 \leq & nSE_n \left[\left(\rho_\tau\left(\eta_s^* - \frac{\theta' Z_s}{\sqrt{n}\sqrt{Sh^k}}\right) - \rho_\tau\left(\eta_s^*\right)\right) - \frac{1}{\sqrt{n}\sqrt{Sh^k}} \rho_\tau'\left(\eta_s^*\right) \theta' Z_s\right]^2 \kappa_s^2 \\ \leq & nSE_n \left[1\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}} \theta' Z_s\right) \left(\frac{1}{\sqrt{n}\sqrt{Sh^k}} \theta' Z_s\right)^2\right] \kappa_s^2 \\ \leq & E_n \frac{1}{h^k} \left(\theta' Z_s\right)^2 \kappa_s^2 P_n\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}} \theta' Z_s | y_s\right). \end{aligned} \tag{33}$$

Since θ is in a compact set and $\kappa(\cdot)$ has bounded support M, $\theta' Z_s \lesssim M$. Also note that

$$\eta_s^* = \eta_s - \eta\left(\check{\theta}_{y_s}\right) + \eta\left(\check{\theta}_{y_s}\right) - \bar{\eta}_\tau^{y_s} + \bar{\eta}_\tau^{y_s} - a_0 - b_0 y_s = \eta_s - \eta\left(\check{\theta}_{y_s}\right) + \frac{1}{\sqrt{n}} q_\tau \sqrt{\Sigma_{\eta(y)}} + o_P\left(\frac{1}{\sqrt{n}}\right),$$

due to (67), (66) and (68). Next by (65) and (73), with $Sh^k \to \infty$, $|\theta| / \sqrt{Sh^k} = o(1)$,

$$\begin{aligned} & P_n\left(\sqrt{n}|\eta_s^*| \leq \frac{1}{\sqrt{Sh^k}} \theta' Z_s | y_s\right) \\ = & \Phi\left(\frac{1}{\sqrt{Sh^k}} \frac{\theta' Z_s}{\sqrt{\Sigma_{\eta(y)}}} - q_\tau + o_P(1)\right) - \Phi\left(-\frac{1}{\sqrt{Sh^k}} \frac{\theta' Z_s}{\Sigma_{\eta(y)}} - q_\tau + o_P(1)\right) + o_P(1) = o_P(1). \end{aligned} \tag{34}$$

This allows us to further bound

$$E_n R_S(\theta)^2 = o_P(1)\, E_n \frac{1}{h^k} \left(\theta' Z_s\right)^2 \kappa_s^2 = o_P(1) \int (1\ v)' \kappa^2(v) f_{Y_n}(vh)\, dv + o_P(1) = o_P(1).$$

Furthermore, even for θ outside a compact set, the same calculation as above shows that for $c_n \to \infty$ but $c_n/\sqrt{Sh^k} \to 0$, $E_n \sup_{|\theta|\leq c_n} R_S(\theta)^2 \leq O_P\left(\frac{c_n^3}{\sqrt{Sh^k}}\right) = o_P\left(c_n^2\right)$, so that

$$\sup_{|\theta|\leq c_n} R_S(\theta) = O_P^*\left(\frac{c_n^{3/2}}{(Sh^k)^{1/4}}\right) = o_P^*(c_n). \tag{35}$$

For example, if $c_n = o_P\left(\sqrt{Sh^k}\right)$, then $\sup_{|\theta|\leq c_n} R_S(\theta) = o_P^*\left(\sqrt{Sh^k}\right)$. See for example Thm 2.14.1 in Van der Vaart and Wellner (1996).

Next, if we can show that

$$W_s \equiv \left(Sh^k\right)^{-1/2} \sum_{s=1}^{S} \rho_\tau'(\eta_s^*) Z_s' \kappa_s = o_P^*\left(\sqrt{Sh^k}\right) \tag{36}$$

then by the same convexity arguments in Fan et al. (1994) and Pollard (1991), we will have $\hat{\theta} + \hat{H}_n^{-1} W_S = o_P^*\left(\sqrt{Sh^k}\right)$, so that $\hat{\theta} = o_P^*\left(\sqrt{Sh^k}\right)$ and thus $\sqrt{n}(\hat{a} - a_0) = o_P^*(1)$. In particular, for v a unit vector, let $B(n)$ be a ball with center $-\hat{H}_n^{-1} W_S$ and radius $\delta_n = o\left(\sqrt{Sh^k}\right)$ but $\hat{H}_n^{-1} W_S = O_P^*(\delta_n)$. For any θ outside $B(n)$, write $\theta = -\hat{H}_n^{-1} W_S + \beta_n v$. Define θ^* as the boundary point of $B(n)$ between $-\hat{H}_n^{-1} W_S$ and θ: $\theta^* = \left(1 - \frac{\delta_n}{\beta_n}\right)\left(-\hat{H}_n^{-1} W_S\right) + \frac{\delta_n}{\beta_n}\theta$. Also let $\Delta_n = \sup_{\theta \in B(n)} R_S(\theta)$. By convexity, for $\eta_n = -\hat{H}_n^{-1} W_S$,

$$\begin{aligned}
\frac{\delta_n}{\beta_n} G_S(\theta) + \left(1 - \frac{\delta_n}{\beta_n}\right) G_S(\eta_n) &\geq G_S(\theta^*) \geq \delta_n' v' \hat{H}_n v \delta_n - \eta_n' \hat{H}_n \eta_n - \Delta_n \\
&\geq \delta_n' v' \hat{H}_n v \delta_n + G_S(\eta_n) - 2\Delta_n
\end{aligned}$$

This leads to $\inf_{|\theta - \eta_n| \geq \delta_n} G_S(\theta) \geq G_S(\eta_n) + \frac{\beta_n}{\delta_n}\left(\delta_n v' \hat{H}_n v \delta_n - 2\Delta_n\right)$. By (35), $\Delta_n = o_P^*(\delta_n)$, so that $P_n\left(\delta_n v' \hat{H}_n v \delta_n - 2\Delta_n \geq 0\right) = 1 - o_P(1)$ when δ_n is bounded away from zero. Namely,

$$P_n\left(\inf_{|\theta - \eta_n| \geq \delta_n} G_S(\theta) \geq G_S(\eta_n)\right) = 1 - o_P(1) \Longrightarrow P_n\left(|\hat{\theta} - \eta_n| \leq \delta_n\right) = 1 - o_P(1).$$

So we can declare $\hat{\theta} = o_P^*\left(\sqrt{Sh^k}\right)$ since both η_n and δ_n are $o_P^*\left(\sqrt{Sh^k}\right)$.

To verify (36), we check both $Var_n(W_S)$ and $E_n(W_S)$.

$$\begin{aligned}
Var_n(W_S) =& \frac{1}{h^k} Var_n\left(\rho_\tau'(\eta_s^*) Z_s' \kappa_s\right) \\
=& \frac{1}{h^k}\left[E_n Var_n\left(\rho_\tau'(\eta_s^*) Z_s' \kappa_s | y_s\right) + Var_n E_n\left(\rho_\tau'(\eta_s^*) Z_s' \kappa_s | y_s\right)\right].
\end{aligned} \tag{37}$$

Recalling $\rho_\tau'(\eta_s^*) = \tau - 1(\eta_s - a_0 - b_0 y_s \leq 0)$, it can be calculated that

$$\begin{aligned}
&E_n Var_n\left(\rho_\tau'(\eta_s^*) Z_s' \kappa_s | y_s\right) = E_n Z_s Z_s' \kappa_s^2 Var_n\left(\tau - 1(\eta_s \leq a_0 + b_0 y_s) | y_s\right) \\
&= \int Z_s Z_s \kappa_s^2 P_n(\eta_s \leq a_0 + b_0 y_s | y_s)\left(1 - P_n(\eta_s \leq a_0 + b_0 y_s | y_s)\right) f_{Y_n}(y_s)\, dy_s.
\end{aligned}$$

Again using (67), (66) and (68), $P_n\left(\eta_s^* \leq 0|y_s\right) = P_n\left(\sqrt{n}\eta_s^* \leq 0|y_s\right) = \tau + o_P(1)$. Hence by the usual change of variable $y_s = vh$, for $\bar{C}_\kappa = \int (1\ v)(1\ v)' \kappa^2(v)\, dv$,

$$\begin{aligned}\frac{1}{h^k} E_n Var_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) =& \frac{1}{h^k} \int Z_s Z_s \kappa_s^2 \left(\tau(1-\tau) + o_P(1)\right) f_{Y_n}(y_s)\, dy_s \\ =& \tau(1-\tau)\, \bar{C}_\kappa f_{Y_n}^\infty(0) + o_P(1).\end{aligned}$$

Next, also using $P_n\left(\sqrt{n}\eta_s^* \leq 0|y_s\right) = \tau + o_P(1)$,

$$\begin{aligned}&\frac{1}{h^k} Var_n E_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s' \kappa_s | y_s\right) = \frac{1}{h^k} Var_n\left[Z_s \kappa_s\left(\tau - P_n\left(\sqrt{n}\eta_s^* \leq 0|y_s\right)\right)\right] \\ &= \frac{1}{h^k} Var_n\left[Z_s \kappa_s o_P(1)\right] \leq \frac{1}{h^k} E_n\left[Z_s \kappa_s o_P(1)\right]^2 = \tilde{C}_\kappa f_{Y_n}^\infty(0)\, o_P(1) = o_P(1).\end{aligned}$$

where $\tilde{C}_\kappa = \int (1\ v)(1\ v)' v^4 \kappa^2(v)\, dv$. Therefore there is $Var_n(W_s) = O_P(1)$.

Consider finally the bias term:

$$\begin{aligned}E_n W_s =& \sqrt{Sh^k} E_n \frac{1}{h^k} E_n\left(\tau - 1\left(\eta_s \leq a_0 + b_0 y_s\right)|y_s\right) Z_s \kappa_s = \sqrt{Sh^k} E_n \frac{1}{h^k}\left(\tau - P_n\left(\eta_s^* \leq 0|y_s\right)\right) Z_s \kappa_s \\ =& \sqrt{Sh^k} o_P(1) \int (1\ v)' \kappa(v) f_{Y_n}(vh)\, dv = \sqrt{Sh^k} o_P(1) f_{Y_n}^\infty(0) \int (1\ v)' \kappa(v)\, dv = o_P\left(\sqrt{Sh^k}\right).\end{aligned} \tag{38}$$

This together with $Var_n(W_s) = O_P(1)$ implies $W_S = o_P^*\left(\sqrt{Sh^k}\right)$. Given these results, the feasible $\hat{\eta}_\tau$ provide valid inference whenever the infeasible posterior quantiles $\bar{\eta}_\tau$ are valid:

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

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

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

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

Overidentification in local linear quantile regression Consider the same change of variable as in the mean regression case in (23). Also for $b_0 = \eta'(y = 0)$, define a similar reparameterization analogous to the mean case:

$$d_1 = \left(b_1 + c' b_2\right), \quad d_2 = b_2/\sqrt{n}, \quad \eta_s^* = \eta_s - a_0 - b_0' y_s = \eta_s - a_0 - d_0' w_s.$$

Also let $\kappa_s = \kappa\left(\frac{y_s}{h}\right) = \kappa\left(\frac{w_{1s}}{h}, \frac{cw_{1s}}{h} + \frac{w_{2s}}{\sqrt{n}h}\right)$ and $Z_s = \left(1, \frac{w_1^s}{h}, w_2^s\right)$. Define

$$\theta = \sqrt{n}\sqrt{Sh^k}\left(\hat{a} - a_0, h\left(\hat{d}_1 - d_{10}\right), \left(\hat{d}_2 - d_{20}\right)\right).$$

Then with this definition $\hat{\theta}$ minimizes the same Koenker and Bassett (1978) check function in (27), which admits the decomposition in (28) and (29), in which $\frac{E_n(\rho''(\eta_s^*)|y_s)}{\sqrt{n}}$ also satisfies the relation in (30). Next define $\bar{\phi}_\tau(y) = \phi\left(q_\tau\sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta\left(\theta_{y_s}\right) - \eta\left(\check{\theta}_{y_s}\right)\right); \Sigma_{\eta(y)}\right)$. Then by changing variables $y = (y_1,\ y_2) = \left(w_1,\ cw_1 + \frac{w_2}{\sqrt{n}}\right)$, and further $w_1 = uh$, we can write, similar to (26)

$$E_n\hat{H}_n = \frac{1}{Sh^k}\sum_{s=1}^{S}\frac{E_n\left(\rho_\tau''\left(\eta_s^*\right)|y_s\right)}{\sqrt{n}}Z_sZ_s'\kappa_s = \int\frac{1}{h^k}\frac{E_n\left(\rho_\tau''\left(\eta_s^*\right)|y_s = \left(w_1,\ cw_1 + \frac{w_2}{\sqrt{n}}\right)\right)}{\sqrt{n}}Z_sZ_s'\kappa_s f_W(w)\,dw$$

$$= \frac{1}{h^k}\int\left[\bar{\phi}_\tau\left(w_1,\ cw_1 + \frac{w_2}{\sqrt{n}}\right) + o_P(1)\right]\begin{pmatrix}1\\ \frac{w_1}{h}\\ w_2\end{pmatrix}\left(1\ \frac{w_1}{h}\ w_2\right)\kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right)\left(f_W^\infty(w) + o_P(1)\right)dw$$

$$= \left(1 + o_P(1)\right)\bar{\phi}_\tau(0)\,\bar{C}_0\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}(1\ u\ w_2)\,\kappa(u, cu)\,e^{-(w_2 - O_P(1))'C_{22}(w_2 - O_P(1)) + O_P(1)}dudw_2$$

$$= \left(1 + o_P(1)\right)H + o_P(1) \quad \text{for} \quad H \quad \text{positive definite w.p} \to 1.$$

Similar calculations as in the exact identification case can verify that both (31) and (32) continue to hold. We also have (34), for $|\theta|/\sqrt{Sh^k} = o(1)$: $Z_s = O_P(1)$, and

$$P_n\left(|\sqrt{n}\eta_s^*| \le \frac{1}{\sqrt{Sh^k}}\theta' Z_s | y_s = \left(w_{1s}, cw_{1s} + \frac{w_{2s}}{\sqrt{n}h}\right)\right) = o_P(1).$$

Next we show (35) for $|c_n| = o\left(\sqrt{Sh^k}\right)$. As before $E_nR_S(\theta) = 0$, so we bound the second moment of an envelope similar to (33):

$$E\sup_{|\theta|\le c_n}R_S(\theta)^2 \le nSE_n\left[1\left(|\sqrt{n}\eta_s^*| \le \frac{1}{\sqrt{Sh^k}}c_n|Z_s|\right)\left(\frac{1}{\sqrt{n}\sqrt{Sh^k}}c_n|Z_s|\right)^2\right]\kappa_s^2$$

$$\le E\frac{1}{h^k}\left(c_n|Z_s|\right)^2\kappa_s^2 P\left(|\sqrt{n}\eta_s^*| \le \frac{1}{\sqrt{Sh^k}}c_n'|Z_s|\middle| y_s = \left(w_{1s}, cw_{1s} + \frac{w_{2s}}{\sqrt{n}}\right)\right)$$

Using $w_1 = O_P^*(h)$, $\frac{w_2}{\sqrt{n}} = O_P^*(h)$, so $P\left(|\sqrt{n}\eta_s^*| \le \frac{1}{\sqrt{Sh^k}}c_n'|Z_s|\middle| y_s = \left(w_{1s}, cw_{1s} + \frac{w_{2s}}{\sqrt{n}}\right)\right) = o_P(1)$,

$$E_n\sup_{|\theta|\le c_n}R_S(\theta)^2 \le o_P(1)\,E_n\frac{1}{h^k}\left(c_n|Z_s|\right)^2\kappa_s^2 \le c_n^2 o_P(1)\,E_n\frac{1}{h^k}|Z_s|^2\kappa_s^2, \quad \text{where}$$

$$E_n\frac{1}{h^k}|Z_s|^2\kappa_s^2 \leq \frac{1}{h^k}\int\left(1+\left(\frac{w_1}{h}\right)^2+w_2^2\right)\kappa^2\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)\left(f_W^\infty(w)+o_P(1)\right)dw$$
$$=\int\left(1+u^2+w_2^2\right)\kappa^2(u,cu)\,e^{-(w_2-O_P(1))'C_{22}(w_2-O_P(1))+O_P(1)}dudw_2+o_P(1)=O_P(1).$$

Therefore $\sup_{|\theta|\leq c_n}R_S(\theta)=o_P^*(c_n)=o_P^*\left(\sqrt{Sh^k}\right)$.

The last step is to show (36) to allow for the same remaining arguments in the exactly identified case. For this purpose we again check both $Var_n(W_S)$ and $E_n(W_S)$,

$$Var_n(W_S)=\frac{1}{h^k}Var_n\left(\rho_\tau'(\eta_s^*)Z_s'\kappa_s\right)$$
$$=\frac{1}{h^k}\left[E_nVar_n\left(\rho_\tau'(\eta_s^*)Z_s'\kappa_s|y_s=\left(w_1,cw_1+\frac{w_2}{\sqrt{n}}\right)\right)+Var_nE_n\left(\rho_\tau'(\eta_s^*)Z_s'\kappa_s|y_s=\left(w_1,cw_1+\frac{w_2}{\sqrt{n}}\right)\right)\right].$$

Recall that $\rho_\tau'(\eta_s^*)=\tau-1(\eta_s\leq a_0+b_0y_s)$, for $y(w)=\left(w_1,cw_1+\frac{w_2}{\sqrt{n}}\right)$,:

$$\frac{1}{h^k}E_nVar_n\left(\rho_\tau'(\eta_s^*)Z_s'\kappa_s|y_s\right)=\frac{1}{h^k}E_nZ_sZ_s'\kappa_s^2Var_n\left(\tau-1(\eta_s\leq a_0+b_0y_s)|y_s=\left(w_{1s},cw_{1s}+\frac{w_{2s}}{\sqrt{n}}\right)\right)$$
$$=\frac{1}{h^k}\int Z_sZ_s\kappa_s^2P_n(\eta_s\leq a_0+b_0y_s|y_s=y(w_s))(1-P_n(\eta_s\leq a_0+b_0y_s|y_s=y(w_s)))f(w_s)dw_s$$
$$=(\tau(1-\tau)+o_P(1))\frac{1}{h^k}\int\begin{pmatrix}1\\ \frac{w_1}{h}\\ w_2\end{pmatrix}\begin{pmatrix}1 & \frac{w_1}{h} & w_2\end{pmatrix}\kappa^2\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)f_w(w_1,w_2)dw_1dw_2$$
$$=(\tau(1-\tau)+o_P(1))\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\begin{pmatrix}1 & u & w_2\end{pmatrix}\kappa^2\left(u,cu+\frac{w_2}{\sqrt{n}h}\right)f_w^\infty(uh,w_2)dudw_2$$
$$=(1+o_P(1))H+o_P(1)\quad\text{for}\quad H\quad\text{positive definite w.p}\to 1,$$

noting that $f_w^\infty(uh,w_2)=\bar{C}_0e^{-(w_2-O_P(1))'C_{22}(w_2-O_P(1))+O_P(1)}$.

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

$$\frac{1}{h^k}Var_nE_n\left(\rho_\tau'(\eta_s^*)Z_s'\kappa_s|y_s=y(w_s)\right)=Var_nZ_s\kappa_s(\tau-P_n(\eta_s\leq a_0+b_0y_s|y_s=y(w_s)))$$
$$=\frac{1}{h^k}Var_n(Z_s\kappa_so_P(1))\leq o_P(1)\frac{1}{h^k}E_nZ_sZ_s'\kappa_s^2$$
$$=o_P(1)\frac{1}{h^k}\int\begin{pmatrix}1\\ \frac{w_1}{h}\\ w_2\end{pmatrix}\begin{pmatrix}1 & \frac{w_1}{h} & w_2\end{pmatrix}\kappa^2\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)f_w(w_1,w_2)dudw_2$$
$$=o_P(1)\int\begin{pmatrix}1\\ u\\ w_2\end{pmatrix}\begin{pmatrix}1 & u & w_2\end{pmatrix}\kappa^2\left(u,cu+\frac{w_2}{\sqrt{n}h}\right)f_w^\infty(uh,w_2)dudw_2=o_P(1)$$

Therefore there is $Var_n(W_S) = O_P(1)$. Finally, consider the bias term, for $\bar{C} = O_P(1)$,

$$\begin{aligned}
E_n W_s &= \sqrt{Sh^k} E_n \frac{1}{h^k} \left(\tau - P_n\left(\eta_s \leq a_0 + b_0 y_s\right) | y_s = y\left(w_s\right)\right) Z_s \kappa_s \\
&= o_P(1) \sqrt{Sh^k} E_n \frac{1}{h^k} Z_s \kappa_s \\
&= o_P(1) \sqrt{Sh^k} \int \frac{1}{h^k} \begin{pmatrix} 1 \\ \frac{w_1}{h} \\ w_2 \end{pmatrix} \kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) f_w\left(w_1, w_2\right) dw_1 dw_2 \\
&= o_P(1) \sqrt{Sh^k} \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \kappa\left(u, cu + o(1)\right) \bar{C} e^{-(w_2 - O_P(1))' C_{22} (w_2 - O_P(1)) + O_P(1)} du dw_2 \\
&= o_P(1) \sqrt{Sh^k} \int \begin{pmatrix} 1 \\ u \\ w_2 \end{pmatrix} \kappa\left(u, cu\right) \left(f_w^{\infty}\left(uh, w_2\right) + o_P(1)\right) du dw_2 = o_P\left(\sqrt{Sh^k}\right).
\end{aligned}$$

Therefore $W_S = o_P^*\left(\sqrt{Sh^k}\right)$ and the same arguments from the exactly identified case apply to verify that $\hat{\eta}_\tau - \bar{\eta}_\tau = \hat{a} - a_0 = o_P^*\left(\frac{1}{\sqrt{n}}\right)$, so that the feasible quantiles $\hat{\eta}_\tau$ provides asymptotically valid confidence intervals whenever $\bar{\eta}_\tau$ does.

B Technical Addendum

B.1 Proofs of theorems 2 and 3

Proof of Theorem 2

Exact Identification We consider first the exact identification case. The arguments for the overidentified case is similar to local linear regressions with properly defined notations. Define $b_u = h^{[u]}\left(\beta_u - \beta_u^0\right)$, 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 $\eta_s^* = \eta_s - \beta_0' y_s^A = \eta_s - b_0' Z_s^A$.

Mean regression We can now write

$$\sqrt{n}\left(\hat{b} - b_0\right) = \left(\frac{1}{Sh^k} \sum_{s=1}^{S} Z_s^A Z_s^A \kappa_s\right)^{-1} \left(\frac{\sqrt{n}}{Sh^k} \sum_{s=1}^{S} Z_s^A \eta_s^* \kappa_s\right) = H^{-1} J.$$

Consider H first, recall that $f_{Y_n}(y) = f_{Y_n}^{\infty}(y)(1+o_P(1))$ Then for

$$C_\kappa = \int v_A v_A' \kappa(v)\, dv, \qquad v_A = \left(v^u = v_1^{u_1} \dots v_d^{u_d}, u \in A\right),$$

$$E_n H = \frac{1}{h^k}\int Z_s^A Z_s^A \kappa_s f_{Y_n}(y^s)\, dy^s = (1+o_P(1))\int v_A v_A' \kappa(v) f_{Y_n}^{\infty}(vh)\, dv$$

$$= (1+o_P(1)) f_{Y_n}^{\infty}(0) C_\kappa + o_P(1) \quad \text{where} \quad f_{Y_n}^{\infty}(0)^{-1} = O_P(1).$$

The variance of a typical element of H takes the form of, for each $[u],[w] \leq p$,

$$Var_n\left(\sqrt{Sh^k}H\right) = \frac{1}{h^k} Var_n\left(Z_s^u Z_s^w \kappa_s\right) = E_n \frac{1}{h^k}(Z_s^u)^2 (Z_s^w)^2 \kappa_s^2 - h^k \left(E_n H_{u,w}\right)^2$$

$$= \int (v^u)^2 (v^w)^2 \kappa(v)^2 f_{Y_n}(vh)\, dv - h^k \left(E_n H_{u,w}\right)^2$$

$$= (1+o_P(1)) f_{Y_n}^{\infty}(0) \int (v^u)^2 (v^w)^2 \kappa^2(v)\, dv + o_P(1).$$

Hence as before $Var_n(H) = o_P(1)$ and $H = f_{Y_n}^{\infty}(0) C_\kappa + o_P^*(1)$.

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

$$E_n J = \frac{\sqrt{n}}{h^k} E_n Z_s^A \kappa_s \left(E(\eta|y^s) - \beta_0' y_s^A\right)$$

$$= \frac{\sqrt{n}}{h^k} E_n Z_s^A \kappa_s \left(\sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta(0) + O\left(|y_s|^{p+2}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right)$$

$$= \sqrt{n}\int v^A \kappa(v) f_{Y_n}(vh) \left(h^{p+1}\sum_{[u]=p+1} v^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta(0) + O\left((vh)^{p+2}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) dv$$

$$= (1+o_P(1)) f_{Y_n}^{\infty}(0) \left[\sqrt{n}h^{p+1}\int v^A \kappa(v) \sum_{[u]=p+1} v^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta(0)\, dv + o_P(1)\int v^A \kappa(v)\, dv\right],$$

so that $E_n J = o_P(1)$ since $\sqrt{n}h^{p+1} \to 0$. Next consider the variance. Note that

$$Var_n\left(\sqrt{Sh^k}J\right) = \frac{n}{h^k} Var_n\left(Z_s^A \kappa_s \eta_s^*\right) = \frac{n}{h^k}\left[E_n Var_n\left(Z_s^A \kappa_s \eta_s^* | y_s\right) + Var_n E_n\left(Z_s^A \kappa_s \eta_s^* | y_s\right)\right].$$

For the first term, by (65) $Var_n(\sqrt{n}\eta|y^s) = J(y^s)^{-1} + o_P(1)$ uniformly in $y \in \mathcal{Y}$,

$$\frac{n}{h^k} E_n Var_n\left(Z_s^A \kappa_s \eta_s^* | y_s\right) = \frac{1}{h^k} E_n Z_s^A Z_s^{A'} \kappa_s^2 Var_n\left(\sqrt{n}\eta^s | y^s\right)$$

$$= \frac{1}{h^k}\int Z_s^A Z_s^{A'} \kappa_s^2 \left[J(y_s)^{-1} + o_P(1)\right] f_{Y_n}(y_s)\, dy_s$$

$$= (1+o_P(1)) J(0)^{-1} f_{Y_n}^{\infty}(0) \int v_A v_A' \kappa^2(v)\, dv + o_P(1).$$

For the second term,

$$\frac{n}{h^k} Var_n Z_s^A \kappa_s E_n\left(\eta_s^* | y_s\right)$$

$$= \frac{n}{h^k} Var_n Z_s^A \kappa_s \left[o_P\left(\frac{1}{\sqrt{n}}\right) + O\left(|y|^{p+2}\right) + \sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta(0)\right] \tag{39}$$

This can be bounded from above by

$$\begin{aligned}
&\frac{n}{h^k}E_n Z_s^A Z_s^{A'}\kappa_s^2\left[o_P\left(\frac{1}{\sqrt{n}}\right)+O\left(|y_s|^{p+2}\right)+\sum_{[u]=p+1}y_s^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)\right]^2\\
\leq&\frac{n}{h^k}E_n Z_s^A Z_s^{A'}\kappa_s^2\left[o_P\left(\frac{1}{n}\right)+O\left(|y_s|^{2(p+2)}\right)+\left(\sum_{[u]=p+1}y_s^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)\right)^2\right]\\
=&\int v_A v_A'\kappa^2\left(v\right)\left(nh^{2(p+1)}\left(\sum_{[u]=p+1}v^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)\right)^2\right.\\
&\left.+o_P\left(1\right)+O\left(nh^{2(p+2)}\right)|v|^{2(p+2)}\right)f_{Y_n}\left(vh\right)dv\\
=&\left(1+o_P\left(1\right)\right)f_{Y_n}^{\infty}\left(0\right)\left[nh^{2(p+1)}\int v_A v_A'\kappa^2\left(v\right)\left(\sum_{[u]=p+1}v^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)\right)^2dv\right.\\
&\left.+o_P\left(1\right)\int v_A v_A'\kappa^2\left(v\right)dv+nh^{2(p+2)}\int v_A v_A'\kappa^2\left(v\right)|v|^{2(p+2)}dv\right]=o_P\left(1\right)
\end{aligned}$$

Since $Sh^k\to\infty$, we conclude that both $Var_n J=o_P\left(1\right)$ and $E_n J=o_P\left(1\right)$. Therefore $J=o_P^*\left(1\right)$. Essentially, we have shown that $J=O_P^*\left(\frac{1}{\sqrt{Sh^k}}\left(1+\sqrt{n}h^{p+1}\right)+\sqrt{n}h^{p+1}\right)$. Then we can write $H^{-1}J=\left(C_\kappa^{-1}\left(f_{Y_n}^{\infty}\left(0\right)\right)^{-1}+o_P^*\left(1\right)\right)J=o_P^*\left(1\right)$.

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 $\eta_s^*=\eta_s-\beta_0' y_s^A$, $\eta_s-\beta' y_s^A=\eta_s^*-b' Z_s^A$. Then $\hat{\theta}$ minimizes

$$G_S\left(\theta\right)=\sqrt{n}\sum_{s=1}^{S}\left(\rho_\tau\left(\eta_s^*-\frac{\theta' Z_s^A}{\sqrt{n}\sqrt{Sh^k}}\right)-\rho_\tau\left(\eta_s^*\right)\right)\kappa_s. \tag{40}$$

Consider now the decomposition, for $\rho_\tau'\left(\cdot\right)=\tau-1\left(\cdot\leq 0\right)$,

$$G_S\left(\theta\right)=E_n\left(G_S\left(\theta\right)|Y\right)+\left(Sh^k\right)^{-1/2}\sum_{s=1}^{S}\left(\rho_\tau'\left(\eta_s^*\right)Z_s^{A'}\kappa_s-E_n\left(\rho_\tau'\left(\eta_s^*\right)|y_s\right)Z_s^{A'}\kappa_s\right)\theta+R_S\left(\theta\right). \tag{41}$$

For bounded support kernel functions, by eq (69),

$$\begin{aligned}
Q_\tau\left(\eta_s^*|y_s\right)=&Q_\tau\left(\eta_s|y_s\right)-\beta_0 y_s^A=\bar{\eta}_\tau^{y_s}-\bar{\eta}_\tau^{y=0}-\beta_{0,-0}' y_s^{A,-0}\\
=&\eta_y-\eta_0-\beta_{0,-0}' y_s^{A,-0}+o_P\left(\frac{1}{\sqrt{n}}\right)=O\left(|y_s|^{p+1}\right)+o_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{42}$$

In the above β_{-0} is the vector of β except the 0th element. Likewise for $y_s^{A,-0}$. We also note

$$\sum_{[u]=p+1} y_s^u \frac{1}{(p+1)!} \frac{\partial^{p+1}}{\partial y^u} \eta(0) = O\left(|y_s|^{p+1}\right). \tag{43}$$

Similar to Eq (29), we write, for $\hat{H}_n \equiv \frac{1}{Sh^k} \sum_{s=1}^S \frac{E_n(\rho_\tau''(\eta_s^*)|y_s)}{\sqrt{n}} Z_s^A Z_s^{A'} \kappa_s$,

$$E_n\left(G_S(\theta)|Y\right) = \left(Sh^k\right)^{-1/2} \sum_{s=1}^S E\left(\rho_\tau'(\eta_s^*)|y_s\right) Z_s^{A'} \kappa_s \theta + (1+o_P(1)) \frac{1}{2} \theta' \hat{H}_n \theta. \tag{44}$$

In the above,

$$E_n\left(\rho_\tau''(\eta_s^*)|y_s\right) = f_{\eta_s^*}^n(0|y_s) = f_{\eta_s}^n\left(\beta_0^A y_s^A|y_s\right) = \sqrt{n} f_{\sqrt{n}(\eta_s - \eta(\check{\theta}_{y_s}))}^n \left(\sqrt{n}\left(\beta_0^A y_s^A - \eta\left(\check{\theta}_{y_s}\right)\right)|y_s\right), \tag{45}$$

Recall as before that $\sup_{y\in\mathcal{Y}} \sup_s \left| f_{\sqrt{n}(\eta-\eta(\check{\theta}_y))}^n(s) - \phi\left(s; \Sigma_{\eta(y)}\right) \right| = o_P(1)$,

$$\begin{aligned}
\sqrt{n}\left(\beta_0^A y_s^A - \eta\left(\check{\theta}_{y_s}\right)\right) =& \sqrt{n}\left(\bar{\eta}_\tau^0 - \eta(\theta_0)\right) + \sqrt{n}\left(\eta(\theta_0) + \beta_{0,-0} y_s^{A,0} - \eta(\theta_{y_s})\right) + \sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right)\right) \\
=& q_\tau \sqrt{\Sigma_{\eta(0)}} + o_P(1) + \sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right)\right)
\end{aligned}$$

by (67) and $\eta(\theta_0) + \beta_{0,-0} y_s^{A,0} - \eta(\theta_{y_s}) = O\left(h^{p+1}\right)$ and $\sqrt{n} h^{p+1} = o(1)$. Therefore (30) also holds,

$$\begin{aligned}
\frac{E_n\left(\rho_\tau''(\eta_s^*)|y_s\right)}{\sqrt{n}} &= \phi\left(q_\tau \sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right)\right) + o_P(1); \Sigma_{\eta(y)}\right) + o_P(1) \\
&= \phi\left(q_\tau \sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta(\theta_0) - \eta\left(\check{\theta}_0\right)\right) + o_P(1); \Sigma_{\eta(0)}\right) + o_P(1),
\end{aligned} \tag{46}$$

using $\sqrt{n}\left(\eta(\theta_{y_s}) - \eta\left(\check{\theta}_{y_s}\right) - \eta(\theta_0) - \eta\left(\check{\theta}_0\right)\right) = o_P(1)$. Change y to vh, for $C_\kappa = \int v_A v_A' \kappa(v)\, dv$,

$$\begin{aligned}
E_n \hat{H}_n \equiv& \int \frac{1}{h^k} \frac{E_n\left(\rho_\tau''(\eta_s^*)|y_s\right)}{\sqrt{n}} Z_s^A Z_s^{A'} \kappa_s f_{Y_n}(y_s)\, dy_s \\
=& \int \frac{E_n\left(\rho_\tau''(\eta_s^*)|y_s = vh\right)}{\sqrt{n}} v_A v_{A'} \kappa(v) f_{Y_n}(vh)\, dv \\
=& f_{Y_n}^\infty(0) \phi\left(q_\tau \sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta(\theta_0) - \eta\left(\check{\theta}_0\right)\right); \Sigma_{\eta(0)}\right) C_\kappa + o_P(1) \equiv H + o_P(1).
\end{aligned} \tag{47}$$

It is also straightforward to show that $Var_n\left(\hat{H}_n\right) = o_P(1)$, so that $\hat{H}_n = H + o_P^*(1)$. Then for $R_S(\theta)$ defined in (41), $E_n R_S(\theta) = 0$, and similar to (32) and (33),

$$G_S(\theta) = \frac{1}{2} \theta' \left(H + o_P^*(1)\right) \theta + \left(Sh^k\right)^{-1/2} \sum_{s=1}^S \rho_\tau'(\eta_s^*) Z_s^{A'} \kappa_s \theta + R_S(\theta). \tag{48}$$

$$\begin{aligned}
E_n R_S(\theta)^2 \leq& nSE_n \left[1\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}} \theta' Z_s^A\right) \left(\frac{1}{\sqrt{n}\sqrt{Sh^k}} \theta' Z_s^A\right)^2\right] \kappa_s^2 \\
\leq& E_n \frac{1}{h^k} \left(\theta' Z_s^A\right)^2 \kappa_s^2 P_n\left(|\eta_s^*| \leq \frac{1}{\sqrt{n}\sqrt{Sh^k}} \theta' Z_s^A | y_s\right).
\end{aligned} \tag{49}$$

Also note that, with $\bar{\eta}_\tau^{y_s} - \beta_0' y_s^A = o_P\left(\frac{1}{\sqrt{n}}\right) + O\left(h^{p+1}\right) = o_P\left(\frac{1}{\sqrt{n}}\right)$,

$$\eta_s^* = \eta_s - \eta\left(\check{\theta}_{y_s}\right) + \eta\left(\check{\theta}_{y_s}\right) - \bar{\eta}_\tau^{y_s} + \bar{\eta}_\tau^{y_s} - \beta_0 y_s^A = \eta_s - \eta\left(\check{\theta}_{y_s}\right) + \frac{1}{\sqrt{n}} q_\tau \sqrt{\Sigma_{\eta(y)}} + o_P\left(\frac{1}{\sqrt{n}}\right), \quad (50)$$

due to (67), (66) and (68). Next by (65) and (73), with $Sh^k \to \infty$, when $|\theta|/\sqrt{Sh^k} = o(1)$, $|Z_s^A| \lesssim M$, $P_n\left(\sqrt{n}|\eta_s^*| \leq \frac{1}{\sqrt{Sh^k}}\theta' Z_s^A | y_s\right) = o_P(1)$ as in (34). This further bounds

$$E_n R_S(\theta)^2 = o_P(1) E_n \frac{1}{h^k}\left(\theta' Z_s^A\right)^2 \kappa_s^2 = o_P(1)|\theta|^2 \min \text{eig} \int v_A v_A' \kappa^2(v) f_{Y_n}(vh)\, dv = o_P(1)|\theta|^2.$$

Then for $|\theta| = O(c_n)$ and $c_n/\sqrt{Sh^k} \to 0$, $E_n \sup_{|\theta|\leq c_n} R_S(\theta)^2 = o_P\left(c_n^2\right)$ and $\sup_{|\theta|\leq c_n} R_S(\theta) = o_P^*(c_n)$. See for example Thm 2.14.1 in Van der Vaart and Wellner (1996).

Next, if we can show that

$$W_s \equiv \left(Sh^k\right)^{-1/2} \sum_{s=1}^S \rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s = o_P^*\left(\sqrt{Sh^k}\right), \quad (51)$$

then by the same arguments as in the proof of Theorem 1, we can conclude that $\hat{\theta} + \hat{H}_n^{-1} W_S = o_P^*\left(\sqrt{Sh^k}\right)$, $\hat{\theta} = o_P^*\left(\sqrt{Sh^k}\right)$ and thus $\sqrt{n}(\hat{a} - a_0) = o_P^*(1)$.

To verify (36), check both $Var_n(W_S)$ and $E_n(W_S)$.

$$\begin{aligned} Var_n(W_S) =& \frac{1}{h^k} Var_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s\right) \\ =& \frac{1}{h^k}\left[E_n Var_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s | y_s\right) + Var_n E_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s | y_s\right)\right]. \end{aligned} \quad (52)$$

Using $\rho_\tau'\left(\eta_s^*\right) = \tau - 1\left(\eta_s^* \leq 0\right)$, and $\eta_s^* = \eta_s - \beta_0' y_s^A$, it can be calculated that

$$\begin{aligned} E_n Var_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^{A'} \kappa_s | y_s\right) =& E_n Z_s^A Z_s^{A'} \kappa_s^2 Var_n\left(\tau - 1\left(\eta_s^* \leq 0\right) | y_s\right) \\ =& \int Z_s^A Z_s^{A'} \kappa_s^2 P_n\left(\eta_s^* \leq 0 | y_s\right)\left(1 - P_n\left(\eta_s^* \leq 0 | y_s\right)\right) f_{Y_n}(y_s)\, dy_s. \end{aligned} \quad (53)$$

Next using (67), (66) and (68), together with (50) and $\sqrt{n}h^{p+1} = o(1)$,

$$P_n\left(\eta_s^* \leq 0 | y_s\right) = P_n\left(\sqrt{n}\eta_s^* \leq 0 | y_s\right) = \Phi\left(q_\tau + o_P(1)\right) + o_P(1) = \tau + o_P(1). \quad (54)$$

Change variable $y_s = vh$, for $\bar{C}_\kappa = \int v_A v_A' \kappa^2(v)\, dv$,

$$\begin{aligned} \frac{1}{h^k} E_n Var_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^A \kappa_s | y_s\right) =& \frac{1}{h^k} \int Z_s^A Z_s^{A'} \kappa_s^2 \left(\tau(1-\tau) + o_P(1)\right) f_{Y_n}(y_s)\, dy_s \\ =& \tau(1-\tau) \bar{C}_\kappa f_{Y_n}^\infty(0) + o_P(1). \end{aligned}$$

Next, consider the second term in variance

$$\begin{aligned} & \frac{1}{h^k} Var_n E_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^A \kappa_s | y_s\right) = \frac{1}{h^k} Var_n Z_s^A \kappa_s E_n\left(\tau - 1\left(\eta_s^* \leq 0\right) | y_s\right) \\ =& \frac{1}{h^k} Var_n Z_s^A \kappa_s \left(\tau - P_n\left(\eta_s^* \leq 0 | y_s\right)\right) \leq o_P(1) \int \frac{1}{h^k} Z_s^A Z_s^{A'} \kappa_s f_{Y_n}(y)\, dy \\ =& o_P(1) \bar{C}_\kappa f_{Y_n}^\infty(0) + o_P(1) = o_P(1). \end{aligned}$$

Therefore there is $Var_n(W_S) = O_P(1)$. Consider finally the bias term:

$$\begin{aligned} E_n W_s =& \sqrt{Sh^k} E_n \frac{1}{h^k} E_n \left(\tau - 1\left(\eta_s^* \leq 0\right)|y_s\right) Z_s^A \kappa_s = \sqrt{Sh^k} E_n \frac{1}{h^k} \left(\tau - P_n\left(\eta_s^* \leq 0|y_s\right)\right) Z_s^A \kappa_s \\ =& \sqrt{Sh^k} o_P(1) \int v_A \kappa(v) f_{Y_n}(vh)\, dv = \sqrt{Sh^k} o_P(1) f_{Y_n}^{\infty}(0) \int v_A \kappa(v)\, dv = o_P\left(\sqrt{Sh^k}\right). \end{aligned}$$

This together with $Var_n(W_s) = O_P(1)$ implies $W_S = o_P^*\left(\sqrt{Sh^k}\right)$. Given these results, the feasible $\hat{\eta}_\tau$ provide valid inference whenever the infeasible posterior quantiles $\bar{\eta}_\tau$ are valid. The same proof can be also adapted for the local polynomial estimator of the posterior mean, using $l(x) = (x)^2$, $l'(x) = 2x$, $l''(x) = 2$, and a different normalization of the objective function:

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

Over Identification Consider the change of variables defined in (23). Then let $\alpha = (\alpha_u, u \in A)$ be implicitly defined in (for $\phi_{s,t}$ being binomial coefficients),

$$\begin{aligned} \beta' y_s^A =& \sum_{u \in A} \beta_u y_s^u = \sum_{u \in A} \beta_u \prod_{j=1}^{k} w_1^{u_j} \prod_{l=k+1}^{d} \left(c w_1 + \frac{w_2}{\sqrt{n}}\right)^{u_l} \\ =& \sum_{u \in A} \beta_u \prod_{j=1}^{k} w_1^{u_j} \prod_{l=k+1}^{d} \sum_{t+s=u_l} \phi_{s,t} (c w_1)^s \left(\frac{w_2}{\sqrt{n}}\right)^t = \sum_{u=(u_1,u_2) \in A} \alpha_u w_1^{u_1} \left(\frac{w_2}{\sqrt{n}}\right)^{u_2}. \end{aligned}$$

Next define $d = (d_u, u \in A)$ as $d_u = d_{u_1,u_2} = \alpha_u h^{u_1} / \sqrt{n}^{u_2}$, $Z_s^A = (Z_s^u, u \in A)$, $Z_s^u = h^{-u_1} w_{1s}^{u_1} w_{2s}^{u_2}$, so that $\beta' y_s^A = d' w_s^A$. Also let $\eta_s^* = \eta^s - \beta' y_s^A = \eta^s - d' w_s^A$. Then write $\eta^s - d' w_s^A = \eta_s^* - (d - d_0)' w_s^A$.

Mean Regression Write $b = d - d_0$, $b_0 = 0$, and

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

Consider change of variables $v_1 = \frac{w_1}{h}$, $v_2 = w_2$, $v = (v_1, v_2)$, using (24) and (25),

$$\begin{aligned} E_n H =& \int \frac{Z_s^A Z_s^{A'} \kappa_s f_W(w)}{h^k} dw = (1 + o_P(1)) \int v_A v_A' \kappa \left(v_1, c v_1 + \frac{v_2}{\sqrt{n} h}\right) f_W^{\infty}(v_1 h, v_2)\, dv \\ =& (1 + o_P(1)) \int v_A v_A' \kappa \left(v_1, c v_1 + \frac{v_2}{\sqrt{n} h}\right) \bar{C}_{y(v_1 h, v_2)} e^{-(v_2 - O_P(1))' C_{22} (v_2 - O_P(1)) + O_P(1)} dv_1 dv_2 \\ =& (1 + o_P(1)) \bar{C}_0 \int v_A v_A' \kappa(v_1, c v_1)\, e^{-(v_2 - O_P(1))' C_{22} (v_2 - O_P(1)) + O_P(1)} dv_1 dv_2 + o_P(1) \\ =& (1 + o_P(1)) \hat{H}_n + o_P(1) \quad \text{for} \quad \hat{H}_n \quad \text{positive definite w.p} \to 1. \end{aligned}$$

The variance of a typical element of H takes the form of, for each $[u],[w]$,

$$Var_n\left(\sqrt{Sh^k}H\right)=\frac{1}{h^k}Var_n\left(Z_s^uZ_s^w\kappa_s\right)\leq E_n\frac{1}{h^k}(Z_s^u)^2(Z_s^w)^2\kappa_s^2$$

$$=\int\left(v^u\right)^2\left(v^w\right)^2\kappa\left(v_1,cv_1+\frac{v_2}{\sqrt{n}h}\right)^2f_W^\infty\left(v_1h,v_2\right)dv$$

$$=\left(1+o_P\left(1\right)\right)\int\left(v^u\right)^2\left(v^w\right)^2\kappa\left(v_1,cv_1+\frac{v_2}{\sqrt{n}h}\right)^2\bar{C}_{y(v_1h,v_2)}e^{-(v_2-O_P(1))'C_{22}(v_2-O_P(1))+O_P(1)}dv_1dv_2$$

$$=\left(1+o_P\left(1\right)\right)\bar{C}_0\int\left(v^u\right)^2\left(v^w\right)^2\kappa\left(v_1,cv_1\right)^2e^{-(v_2-O_P(1))'C_{22}(v_2-O_P(1))+O_P(1)}dv_1dv_2+o_P\left(1\right)=O_P\left(1\right).$$

Therefore $Var_n\left(H\right)=o_P\left(1\right)$, and $H=\hat{H}_n+o_P^*\left(1\right)$. Now consider the bias of J, for $y=\left(w_1,cw_1+\frac{w_2}{\sqrt{n}}\right)=\left(hv_1,chv_1+\frac{v_2}{\sqrt{n}}\right)$, and for $\bar{v}=\left(v_1,cv_1+\frac{v_2}{\sqrt{n}h}\right)$, and for

$$f_{V_2}^\infty\left(v_2\right)=\bar{C}_0e^{-(v_2-O_P(1))'C_{22}(v_2-O_P(1))+O_P(1)},$$

$$E_nJ=\frac{\sqrt{n}}{h^k}E_nZ_s^A\kappa_s\left(\sum_{[u]=p+1}y_s^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)+O\left(\left|y_s\right|^{p+2}\right)+o_P\left(\frac{1}{\sqrt{n}}\right)\right)$$

$$=\sqrt{n}\int v^A\kappa\left(\bar{v}\right)f_W^\infty\left(v_1h,v_2\right)\left(h^{p+1}\sum_{[u]=p+1}\bar{v}^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)+O\left(\left(\bar{v}h\right)^{p+2}\right)+o_P\left(\frac{1}{\sqrt{n}}\right)\right)dv$$

$$=\left(1+o_P\left(1\right)\right)\left[\sqrt{n}h^{p+1}\int v^A\kappa\left(v_1,cv_1\right)\sum_{[u]=p+1}\left(v_1,cv_1\right)^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)f_{V_2}^\infty\left(v_2\right)dv\right.$$

$$\left.+o_P\left(1\right)\int v^A\kappa\left(v_1,cv_1\right)f_{V_2}^\infty\left(v_2\right)dv+o_P\left(1\right)\right]=o_P\left(1\right).$$

Next consider the variance: $Var_n\left(\sqrt{Sh^k}J\right)=\frac{n}{h^k}\left[E_nVar_n\left(Z_s^A\kappa_s\eta_s^*|y_s\right)+Var_nE_n\left(Z_s^A\kappa_s\eta_s^*|y_s\right)\right]$.

$$\frac{n}{h^k}E_nVar_n\left(Z_s^A\kappa_s\eta_s^*|y_s\right)=\frac{1}{h^k}E_nZ_s^AZ_s^{A'}\kappa_s^2Var_n\left(\sqrt{n}\eta^s|y^s\right)=\frac{1}{h^k}\int Z_s^AZ_s^{A'}\kappa_s^2\left[J\left(y_s\right)^{-1}+o_P\left(1\right)\right]f_W\left(w_s\right)dw_s$$

$$=\left(1+o_P\left(1\right)\right)J\left(0\right)^{-1}\int v_Av_A'\kappa^2\left(v_1,cv_1\right)f_{V_2}^\infty\left(v_2\right)dv+o_P\left(1\right).$$

The second variance term in (39) can be bounded by

$$\int v_Av_A'\kappa^2\left(\bar{v}\right)\left(nh^{2(p+1)}\left(\sum_{[u]=p+1}\bar{v}^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)\right)^2\right.$$

$$\left.+o_P\left(1\right)+O\left(nh^{2(p+2)}\right)\left|\bar{v}\right|^{2(p+2)}\right)f_W\left(v_1h,v_2\right)dv$$

$$=\left(1+o_P\left(1\right)\right)\left[nh^{2(p+1)}\int v_Av_A'\kappa^2\left(v_1,cv_1\right)\left(\sum_{[u]=p+1}\left(v_1,cv_1\right)^u\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial y^u}\eta\left(0\right)\right)^2f_{V_2}^\infty\left(v_2\right)dv\right.$$

$$\left.+o_P\left(1\right)\int v_Av_A'\kappa^2\left(v_1,cv_1\right)f_{V_2}^\infty\left(v_2\right)dv+nh^{2(p+2)}\int v_Av_A'\kappa^2\left(v_1,cv_1\right)\left|\left(v_1,cv_1\right)\right|^{2(p+2)}f_{V_2}^\infty\left(v_2\right)dv\right]=o_P\left(1\right)$$

Therefore $Var_n\left(J\right)=o_P\left(1\right)$, $J=o_P^*\left(1\right)$, and $\sqrt{n}\hat{b}=H^{-1}J=o_P^*\left(1\right)$.

Quantile Regression Let $\theta = \sqrt{n}\sqrt{Sh^k}(d - d_0)$, which minimizes (40). Then (42), (43), (44), (45) and (46) all continue to hold. Let $\phi_0^\tau = \phi\left(q_\tau\sqrt{\Sigma_{\eta(0)}} + \sqrt{n}\left(\eta(\theta_0) - \eta\left(\check{\theta}_0\right)\right); \Sigma_{\eta(0)}\right)$. Change (47) to

$$\begin{aligned} E_n\hat{H}_n &\equiv \int \frac{E_n\left(\rho_\tau''\left(\eta_s^*\right)|y_s = \bar{v}h\right)}{\sqrt{n}} v_A v_{A'}\kappa(\bar{v}) f_W(v_1 h, v_2)\, dv \\ &= (1 + o_P(1))\,\phi_0^\tau \int v_A v_A'\kappa(v_1, cv_1) f_{V_2}^\infty(v_2)\, dv + o_P(1) \equiv H + o_P(1). \end{aligned}$$

It can also be shown $Var_n\left(\hat{H}_n\right) = o_P(1)$, so $H_n = H + o_P^*(1)$. Next (48), (49), (50) all continue to hold. When $Sh^k \to \infty$ and when $|\theta|/\sqrt{Sh^k} = o(1)$, pointwise in Z_s^A, $P_n\left(\sqrt{n}|\eta_s^*| \leq \frac{1}{\sqrt{Sh^k}}\theta' Z_s^A|y_s\right) = o_P(1)$. Then by dominated convergence,

$$\begin{aligned} E_n R_S(\theta)^2 &= o_P(1) E_n \frac{1}{h^k}\left(\theta' Z_s^A\right)^2 \kappa_s^2 = o_P(1)|\theta|^2 \min \text{eig} \int v_A v_A'\kappa^2(\bar{v}) f_W(v_1 h, v_2)\, dv \\ &= o_P(1)|\theta|^2 \min \text{eig}\left[\int v_A v_A'\kappa^2(v_1, cv_1) f_{V_2}^\infty(v_2)\, dv + o_P(1)\right] = o_P(1)|\theta|^2. \end{aligned}$$

Same as before for $|\theta| = O(c_n)$ and $c_n/\sqrt{Sh^k} \to 0$, $E_n \sup_{|\theta|\leq c_n} R_S(\theta)^2 = o_P\left(c_n^2\right)$ and $\sup_{|\theta|\leq c_n} R_S(\theta) = o_P^*(c_n)$. It remains to verify (51) by checking $E_n(W_S)$ and $Var_n(W_S)$ via (52). Next (53) and (54) continue to hold. Then we write the first term in $Var_n(W_S)$ as

$$\begin{aligned} &\frac{1}{h^k} E_n Var_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^A \kappa_s | y_s\right) = \frac{1}{h^k}\int Z_s^A Z_s^{A'}\kappa_s^2\left(\tau(1-\tau) + o_P(1)\right) f_W(w_s)\, dw_s \\ &= (\tau(1-\tau) + o_P(1)) \int v_A v_{A'}\kappa_s^2(\bar{v}) f_W(v_1 h, v_2)\, dv + o_P(1) \\ &= (\tau(1-\tau) + o_P(1)) \int v_A v_{A'}\kappa_s^2(v_1, cv_1) f_{V_2}^\infty(v_2)\, dv + o_P(1) \end{aligned}$$

and the second term of the variance as

$$\begin{aligned} &\frac{1}{h^k} Var_n E_n\left(\rho_\tau'\left(\eta_s^*\right) Z_s^A \kappa_s | y_s\right) \leq o_P(1)\int \frac{1}{h^k} Z_s^A Z_s^{A'}\kappa_s f_W(w_s)\, dw_s \\ =& o_P(1)\int v_A v_{A'}\kappa_s(v_1, cv_1) f_{V_2}^\infty(v_2)\, dv + o_P(1) = o_P(1). \end{aligned}$$

Finally compute the bias

$$\begin{aligned} E_n W_s =& = \sqrt{Sh^k} E_n \frac{1}{h^k}\left(\tau - P_n\left(\eta_s^* \leq 0 | y_s\right)\right) Z_s^A \kappa_s \\ =& \sqrt{Sh^k} o_P(1)\int v_A \kappa(\bar{v}) f_W(v_1 h, v_2)\, dv \\ =& \sqrt{Sh^k} o_P(1)\left[\int v_A \kappa(v_1, cv_1) f_{V_2}(v_2)\, dv + o_P(1)\right] = o_P\left(\sqrt{Sh^k}\right). \end{aligned}$$

The remaining arguments are the same as before.

Proof of Theorem 3 We consider the exact identification case and the overidentification case separately. We prove the mean regression. The proof for quantile regression is similar and omitted.

Exact identification $d = k$. Consider first

$$\sqrt{n}\left(\hat{\eta} - \bar{\eta}_0\right) = \frac{\sqrt{n}A_1 + \sqrt{n}A_2}{A_3}$$

where $A_3 = \frac{1}{Sh^k}\sum_{s=1}^{S}\kappa\left(Y_n^s/h\right)$, $A_2 = E_n\left(\left(\eta^s - \bar{\eta}_0\right)\frac{1}{h^k}\kappa\left(Y_n^s/h\right)\right)$, and

$$A_1 = \frac{1}{Sh^k}\sum_{s=1}^{S}\left(\eta^s - \bar{\eta}_0\right)\kappa\left(Y_n^s/h\right) - E_n\left(\left(\eta^s - \bar{\eta}_0\right)\frac{1}{h^k}\kappa\left(Y_n^s/h\right)\right),$$

Then $E_nA_3 = \int \kappa(v) f_{Y_n}(vh)\,dv = (1 + o_P(1)) f_{Y_n}^{\infty}(0) + o_P(1)$, and

$$\begin{aligned} Var_n(A_3) &= \frac{1}{Sh^k}\frac{1}{h^k}Var_n\kappa\left(\frac{y_n^s}{h}\right) \leq \frac{1}{Sh^k}\int \kappa(v)^2 f_{Y_n}(vh)\,dv \\ &= \frac{1}{Sh^k}\left((1 + o_P(1)) f_{Y_n}^{\infty}(0)\int \kappa^2(v)\,dv + o_P(1)\right) = o_P(1). \end{aligned}$$

Therefore $A_3 = f_{Y_n}^{\infty}(0) + o_P^*(1)$, so that $A_3^{-1} = f_{Y_n}^{\infty}(0)^{-1} + o_P^*(1) = O_P^*(1)$ Next by (68),

$$\begin{aligned} \sqrt{n}A_2 = \frac{\sqrt{n}}{h^k}E_n\kappa_s\left(E_n\left(\eta|y^s\right) - \bar{\eta}_0\right) &= \frac{\sqrt{n}}{h^k}E_n\kappa_s\left(y_s'\eta'(0) + \frac{1}{2}y_s'\eta''(0)\,y_s + O\left(y_s^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\ =\sqrt{n}\int \kappa(v)\left(hv'\eta'(0) + h^2\frac{1}{2}v'\eta''(0)\,v + O\left(v^3h^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) f_{Y_n}(vh)\,dv \\ = (1 + o_P(1)) f_{Y_n}^{\infty}(0)\left[\sqrt{n}h^2\int (1\ v)'\frac{v'\eta''(0)\,v}{2}\kappa(v)\,dv + o_P(1)\right] + o_P(1) = o_P(1). \end{aligned}$$

Since $E_nA_1 = 0$, consider now the conditional variance of A_1,

$$Var_n\left(\frac{\sqrt{n}}{\sqrt{Sh^k}}A_1\right) = \frac{n}{h^k}Var_n\left(\kappa_s\eta_s^*\right) = \frac{n}{h^k}\left[E_nVar_n\left(\kappa_s\eta_s^*|y_s\right) + Var_nE_n\left(\kappa_s\eta_s^*|y_s\right)\right].$$

For the first term, by (65), $Var_n\left(\sqrt{n}\eta|y^s\right) = J\left(y^s\right)^{-1} + o_P(1)$ uniformly in $y \in \mathcal{Y}$,

$$\begin{aligned} &\frac{n}{h^k}E_nVar_n\left(\kappa_s\eta_s^*|y_s\right) = \frac{1}{h^k}E_n\kappa_s^2Var_n\left(\sqrt{n}\eta^s|y^s\right) \\ =&\frac{1}{h^k}\int \kappa_s^2\left[J\left(y_s\right)^{-1} + o_P(1)\right] f_{Y_n}(y_s)\,dy_s = J(0)^{-1} f_{Y_n}^{\infty}(0)\int \kappa^2(v)\,dv + o_P(1). \end{aligned}$$

For the second term,

$$\begin{aligned} &\frac{n}{h^k}Var_n\kappa_sE_n\left(\eta_s^*|y_s\right) = \frac{n}{h^k}Var_n\kappa_s\left(y_s'\eta'(0) + \frac{1}{2}y_s'\eta''(0)\,y_s + O\left(y_s^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\ &= \frac{1}{h^k}E_n\kappa_s^2\left(\sqrt{n}y_s'\eta'(0) + \sqrt{n}\frac{1}{2}y_s'\eta''(0)\,y_s + \sqrt{n}O\left(y_s^3\right) + o_P(1)\right)^2 - h^k\left(E_n\frac{\sqrt{n}}{h^k}\kappa_s\eta_s^*\right)^2 \\ &= nh^2 f_{Y_n}^{\infty}(0)\int \kappa^2(v)\left(v'\eta'(0) + \frac{1}{2}h^2v'\eta''(0)\,v\right)^2 dv + o_P(1). \end{aligned}$$

Therefore, since $Sh^k \min\left(1, nh^2\right) \to \infty$,

$$Var_n\left(\sqrt{n}A_1\right) = O_P\left(\frac{1}{Sh^k} + \frac{nh^2}{Sh^k}\right) = o_P\left(1\right).$$

In the usual situation, the first term dominates the second term in the variation of A_1. In this case however, both terms are important to consider. The stated result follows from combining the above terms so that $\sqrt{n}\left(\hat{\eta} - \bar{\eta}_0\right) = o_P^*\left(1\right)$.

Over identification $d > k$. Consider the change of variables defined in (23). Then $f_W\left(w\right) = \sqrt{n}^{d-k} f_{Y_n}\left(w_1, cw_1 + \frac{w_2}{\sqrt{n}}\right)$ and $\kappa\left(\frac{y}{h}\right) = \kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right)$. Also recall (24) and (25).

$$\begin{aligned}
E_n A_3 &= \int \frac{\kappa_s f_W\left(w\right)}{h^k} dw = \frac{1 + o_P\left(1\right)}{h^k} \int \kappa\left(\frac{w_1}{h}, \frac{cw_1}{h} + \frac{w_2}{\sqrt{n}h}\right) f_W^{\infty}\left(w\right) dw \\
&= \left(1 + o_P\left(1\right)\right) \int \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) \bar{C}_{y(uh, w_2)} e^{-(w_2 - O_P(1))' C_{22} (w_2 - O_P(1)) + O_P(1)} du dw_2 \\
&= \left(1 + o_P\left(1\right)\right) \bar{C}_0 \int \kappa\left(u, cu\right) e^{-(w_2 - O_P(1))' C_{22} (w_2 - O_P(1)) + O_P(1)} du dw_2 + o_P\left(1\right) \\
&= \left(1 + o_P\left(1\right)\right) \hat{H}_n + o_P\left(1\right) \quad \text{for} \quad \hat{H}_n \quad \text{strictly positive w.p} \to 1.
\end{aligned}$$

Similar calculations can also be used to check that $Var_n\left(A_3\right) = o_P\left(1\right)$. Therefore $A_3 = \hat{H}_n + o_P^*\left(1\right)$.

Next, consider the bias term A_2 first. Note that $\eta\left(y\right) = \eta\left(y_1 = w_1, y_2 = cw_1 + \frac{w_2}{\sqrt{n}}\right)$,

$$\begin{aligned}
\sqrt{n}A_2 &= \frac{\sqrt{n}}{h^k} E\kappa_s\left(E\left(\eta | y^s\right) - \bar{\eta}_0\right) = \frac{\sqrt{n}}{h^k} E_n \kappa_s \left(y_s' \eta'\left(0\right) + \frac{1}{2} y_s' \eta''\left(0\right) y_s + O\left(y_s^3\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
= &\sqrt{n} \int \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) \times \left[\begin{pmatrix} uh \\ cuh + \frac{w_2}{\sqrt{n}} \end{pmatrix}' \eta'\left(0\right) + \right. \\
&\left. \frac{1}{2} \begin{pmatrix} uh \\ cuh + \frac{w_2}{\sqrt{n}} \end{pmatrix}' \eta''\left(0\right) \begin{pmatrix} uh \\ cuh + \frac{w_2}{\sqrt{n}} \end{pmatrix} + o\left(u^2 h^2 + \frac{w^2}{n}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right] f_W\left(uh, w_2\right) du dw_2 \\
= &\sqrt{n}h \int \kappa\left(u, cu + o_P\left(1\right)\right) \begin{pmatrix} u \\ cu + o_P\left(1\right) \end{pmatrix}' \eta'\left(0\right) \left(f_W\left(0, w_2\right) + O_P\left(h\right)\right) du dw_2 \\
&+ \sqrt{n}h^2 \int \kappa\left(u, cu + \frac{w_2}{\sqrt{n}h}\right) \frac{1}{2} \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix}' \left(\eta''\left(0\right) + o_P\left(1\right)\right) \begin{pmatrix} u \\ cu + \frac{w_2}{\sqrt{n}h} \end{pmatrix} f_W^{\infty}\left(0, w_2\right) du dw_2 + o_P\left(1\right) \\
= &\sqrt{n}h^2 \int \frac{1}{2} \begin{pmatrix} u \\ cu \end{pmatrix}' \left(\eta''\left(0\right) + o_P\left(1\right)\right) \begin{pmatrix} u \\ cu \end{pmatrix} f_W^{\infty}\left(0, w_2\right) du dw_2 \left(1 + o_P\left(1\right)\right) + o_P\left(1\right)
\end{aligned}$$

The variance also has two terms.

$$Var_n\left(\frac{\sqrt{n}}{\sqrt{Sh^k}} A_1\right) = \frac{n}{h^k} Var_n\left(\kappa_s \eta_s^*\right) = \frac{n}{h^k}\left[E_n Var_n\left(\kappa_s \eta_s^* | y_s\right) + Var_n E_n\left(\kappa_s \eta_s^* | y_s\right)\right].$$

The first in variance

$$\begin{aligned}
&\frac{n}{h^k}E_nVar_n\left(\kappa_s\eta_s^*|y_s\right)=\frac{n}{h^k}E\kappa_s^2Var_n\left(\eta^s|y^s\right)\\
=&\frac{n}{h^k}\int\kappa^2\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)\times Var\left(\eta_s^*|y_1=w_1,y_2=cw_1+\frac{w_2}{\sqrt{n}}\right)f_W\left(w\right)dw\\
=&\int\kappa^2\left(u,cu+\frac{w_2}{\sqrt{n}h}\right)\ n\ Var_n\left(\eta_s^*|y_1=uh,y_2=cuh+\frac{w_2}{\sqrt{n}}\right)f_W\left(uh,w_2\right)dudw_2\\
=&\int\left(\kappa^2\left(u,cu\right)+o_{a.s}\left(1\right)\right)\left(\mathcal{J}^{-1}+o_P\left(1\right)\right)f_W^\infty\left(uh,w_2\right)\left(1+o_P\left(1\right)\right)dudw_2\\
=&\mathcal{J}^{-1}\int\kappa^2\left(u,cu\right)f_W^\infty\left(0,w_2\right)dudw_2+o_P\left(1\right)=O_P\left(1\right).
\end{aligned}$$

The second term in variance,

$$\begin{aligned}
&\frac{n}{h^k}Var_n\kappa_sE\left(\eta_s^*|y_s\right)=\frac{n}{h^k}Var_n\kappa_s\left(y_s'\eta'\left(0\right)+\frac{1}{2}y_s'\eta''\left(0\right)y_s+O\left(y_s^3\right)+o_P\left(\frac{1}{\sqrt{n}}\right)\right)\\
&\leq\frac{1}{h^k}E_n\kappa_s^2\left(\sqrt{n}y_s'\eta'\left(0\right)\sqrt{n}\frac{1}{2}y_s'\eta''\left(0\right)y_s+\sqrt{n}O\left(y_s^3\right)+o_P\left(1\right)\right)^2\\
=&\frac{1}{h^k}\int\kappa^2\left(\frac{w_1}{h},\frac{cw_1}{h}+\frac{w_2}{\sqrt{n}h}\right)\left(\begin{pmatrix}\sqrt{n}w_1\\ c\sqrt{n}w_1+w_2\end{pmatrix}'\eta'\left(0\right)\right.\\
&\left.+\frac{1}{2}\begin{pmatrix}n^{1/4}w_1\\ cn^{1/4}w_1+\frac{w_2}{n^{1/4}}\end{pmatrix}'\eta''\left(0\right)\begin{pmatrix}n^{1/4}w_1\\ cn^{1/4}w_1+\frac{w_2}{n^{1/4}}\end{pmatrix}+\sqrt{n}O\left(w_1^3+\frac{w_2^3}{n\sqrt{n}}\right)+o_P\left(1\right)\right)^2f_W\left(w_1,w_2\right)dw_1dw_2\\
=&nh^2\int\kappa^2\left(u,cu+\frac{w_2}{\sqrt{n}h}\right)\left(\begin{pmatrix}u\\ cu+\frac{w_2}{\sqrt{n}h}\end{pmatrix}'\eta'\left(0\right)\right.\\
&\left.+\frac{1}{2}h\begin{pmatrix}u\\ cu+\frac{w_2}{\sqrt{n}h}\end{pmatrix}'\eta''\left(0\right)\begin{pmatrix}u\\ cu+\frac{w_2}{\sqrt{n}h}\end{pmatrix}+o\left(u^2+\frac{w_2^2}{nh^2}\right)+o_P\left(1\right)\right)^2f_W\left(uh,w_2\right)dudw_2\\
=&nh^2\left[\int\kappa^2\left(u,cu\right)\left(\begin{pmatrix}u\\ cu\end{pmatrix}'\eta'\left(0\right)+\frac{1}{2}h\begin{pmatrix}u\\ cu\end{pmatrix}'\eta''\left(0\right)\begin{pmatrix}u\\ cu\end{pmatrix}\right)^2\left(1+o_P\left(1\right)\right)f_W\left(0,w_2\right)dudw_2+o_P\left(1\right)\right]
\end{aligned}$$

As before, since $Sh^k\min\left(1,nh^2\right)\to\infty$, $Var_n\left(\sqrt{n}A_1\right)=O_P\left(\frac{1}{Sh^k}+\frac{nh^2}{Sh^k}\right)=o_P\left(1\right)$.

B.2 Additional proofs

Proof of Lemma 1 Note that by definition,

$$f_{Y_n}\left(y\right)=\int\pi\left(\theta\right)\sqrt{n}^d\sqrt{2\pi}^{-d/2}\det\left(\hat{\Sigma}\left(\theta\right)\right)^{-1/2}e^{n\hat{Q}_y(\theta)}d\theta,$$

where $\hat{Q}_y\left(\theta\right)$ is either (18) or $\hat{Q}_1^y\left(\theta\right)$ in section B.4. Also define

$$f_{Y_n}^\infty\left(y\right)=\pi\left(\theta_y\right)\sqrt{2\pi}^{-(d-k)/2}\det\left(\Sigma\left(\theta_y\right)\right)^{-1/2}\det\left(J_y\right)^{1/2}e^{n\hat{Q}_y\left(\check{\theta}_y\right)}\equiv\bar{C}_ye^{n\hat{Q}_y\left(\check{\theta}_y\right)}\tag{55}$$

We verify the following stronger statement which implies Lemma 1:

$$\sup_{y\in\mathcal{Y}} |f_{Y_n}(y) / \left(\sqrt{n}^{d-k} f_{Y_n}^{\infty}(y)\right) - 1| = o_P(1). \tag{56}$$

For this purpose, write

$$f_{Y_n}(y)\sqrt{2\pi}^{d/2}\sqrt{n}^{k-d}/e^{n\hat{Q}_y(\check{\theta}_y)} = \sqrt{n}^{k}\int \pi_2(\theta)\, e^{n\left(\hat{Q}_y(\theta)-\hat{Q}_y(\check{\theta}_y)\right)}d\theta = C_n^y,$$

for C_n^y and $\pi_2(\cdot)$ defined in (70) and (64). Then by (72),

$$\sup_{y\in\mathcal{Y}} |f_{Y_n}(y)\sqrt{2\pi}^{d/2}\sqrt{n}^{k-d}/e^{n\hat{Q}_y(\check{\theta}_y)} - \pi_2(\theta_y)(2\pi)^{k/2}\det(J_y)^{-1/2}| = o_P(1). \tag{57}$$

which can be rearranged to (56) as long as $\inf_{y\in\mathcal{Y}} \pi_2(\theta_y)(2\pi)^{k/2}\det(J_y)^{-1/2} > 0$ and

$$\sup_{y\in\mathcal{Y}} |\det\left(\hat{\Sigma}(\theta_y)\right)\det(\Sigma(\theta_y))^{-1} - 1| = o_P(1).$$

B.3 Preliminary Results

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 as intermediate steps in the proof of the theorems in the paper.

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 $\mathcal{Y} \subset \mathbb{R}^d$ is a shrinking neighborhood around zero.

LEMMA 2 Suppose the following three conditions hold. (1) Uniform convergence.

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

(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

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

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

Proof: Condition (3) is automatically satisfied when $\tilde{\theta}(y) = \arg\max \hat{Q}_y(\theta)$. 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 $y \in \mathcal{T}$,

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

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),

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

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. In the following we use $\theta(y)$ and θ_y interchangably.

LEMMA 3 In addition to the conditions in Lemma 2, suppose that

$$\inf_{y\in\mathcal{Y}}\left(\hat{Q}_y\left(\tilde{\theta}_y\right) - \sup_{\theta\in\Theta}\hat{Q}_y(\theta)\right) \geq -o_P\left(n^{-1}\right),$$

and that 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 $\sqrt{n}\Omega^{-1/2}\Delta_n^y \xrightarrow{d} N(0, I)$, 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 $\delta \to 0$

$$\sup_{y\in\mathcal{Y}}\sup_{|\theta-\theta^*|\leq\delta,\theta\in\mathcal{N}(\theta_0),\theta^*\in\mathcal{N}(\theta_0)} \frac{R_n^y(\theta, \theta^*)}{1/n + |\theta - \theta^*|^2 + |\theta - \theta^*|/\sqrt{n}} = o_P(1). \tag{58}$$

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 $\inf_{y\in\mathcal{Y}}\inf_{|x|\neq 0}\frac{x'J_yx}{x'x} > 0$, and for any $\delta_n \to 0$,

$$\sup_{y\in\mathcal{Y}}\sup_{|\theta-\theta_y|\leq\delta_n} \frac{|Q_y(\theta) - Q_y(\theta_y) - \frac{1}{2}(\theta - \theta_y)'H_y(\theta - \theta_y)|}{|\theta - \theta_y|^2} = o(1). \tag{59}$$

Then $\sup_{y\in\mathcal{Y}}|\sqrt{n}\left(\tilde{\theta}_y - \theta_y\right) - J_y^{-1}\sqrt{n}\Delta_n^y| = o_P(1)$.

Proof: We retrace the steps in Newey and McFadden (1994). Note Lemma 2 implies $\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y-\theta_y|=o_P(1)$. First we show that $\sup_{y\in\mathcal{Y}}|\tilde{\theta}_y-\theta_y|=O_P(1)$. By (59), $\exists C>0$ such that for all $y\in\mathcal{Y}$ and all $\theta-\theta_y=o(1)$,

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

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

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

Note that (58) also implies that if we had defined

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

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

$$\sup_{y\in\mathcal{Y}}\sup_{|\theta-\theta^*|\leq\delta,\theta\in\mathcal{N}(\theta_0),\theta^*\in\mathcal{N}(\theta_0)}\frac{\hat{R}^y(\theta,\theta^*)}{|\theta-\theta^*|^2+|\theta-\theta^*|/\sqrt{n}}=o_P(1). \tag{60}$$

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 $y\in\mathcal{Y}$,

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

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 (60), uniformly in $y\in\mathcal{Y}$,

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

and

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

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

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

it follows that

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

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 4 The conditions in Lemmas 2 and 3 hold under Assumptions 1, 2 and one of 3 to 5.

Proof: Verifying the conditions in Lemma 2 is relatively straightforward using Assumption 2, so we focus on those in Lemma 3 using Assumption 2 and one of 3 to 5.

Recall that $Q_y(\theta) = -\frac{1}{2}(g(\theta) - y)' W(\theta)(g(\theta) - y)$, where θ_y is defined by $\frac{\partial}{\partial\theta} Q_y(\theta)\Big|_{\theta=\theta_y} = 0$, where (in the following $\frac{\partial}{\partial\theta} W(\theta)$ is understood elementwise)

$$\frac{\partial}{\partial\theta} Q_y(\theta) = G(\theta)' W(\theta)(g(\theta) - y) + (g(\theta) - y)' \frac{\partial}{\partial\theta} W(\theta)(g(\theta) - y).$$

Furthermore

$$(Q_y(\theta) - Q_y(\theta_y)) = \frac{1}{2}(\theta - \theta_y)' \frac{\partial^2}{\partial\theta\partial\theta'} Q_y(\theta_y)(\theta - \theta_y) + o\left(|\theta - \theta_y|^2\right).$$

But except at $y = 0$, an explicit expression for the above Hessian can be very messy.

This verifies (59). Next we consider the key condition (60). 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

$$\hat{\epsilon} \equiv \hat{\epsilon}(\mathcal{Y}, \delta) = \sup_{y\in\mathcal{Y}, |\theta-\theta_y|\leq\delta} \hat{\epsilon}(\theta, \theta_y) = o_P\left(1/\sqrt{n}\right). \tag{61}$$

Recall that $Q_y(\theta) = -\frac{1}{2}(g(\theta) - y)' W(\theta)(g(\theta) - y)$, and that $\hat{Q}_y(\theta) = -\frac{1}{2}(\hat{g}(\theta) - y)' \hat{W}(\theta)(\hat{g}(\theta) - y)$. By expanding

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

We can decompose

$$\hat{R}^y(\theta,\theta_y) = \hat{Q}_y(\theta) - \hat{Q}_y(\theta_y) - Q_y(\theta) + Q_y(\theta_y) - \Delta_n^{y'}(\theta-\theta_y) = (1)+(2)+(3)+(4)+(5)+(6)$$

where, for $\bar{Q}_y(\theta) = -\frac{1}{2}(g(\theta)-y)' W_y (g(\theta)-y)$

$$\begin{aligned}
(1) = & \quad -\tfrac{1}{2}(g(\theta)-g(\theta_y))'\hat{W}_y(g(\theta)-g(\theta_y)) - \bar{Q}_y(\theta) + Q_y(\theta_y)\\
(2) = & \quad -(1+\sqrt{n}|\theta-\theta_y|)^2\hat{\epsilon}'\hat{W}_y\hat{\epsilon}\\
(3) = & \quad -(\hat{g}(\theta_y)-y)'\hat{W}_y(g(\theta)-g(\theta_y)) + \Delta_n^{y'}(\theta-\theta_y)\\
(4) = & \quad -(g(\theta)-g(\theta_y))'\hat{W}_y\hat{\epsilon}(1+\sqrt{n}|\theta-\theta_y|)\\
(5) = & \quad -(\hat{g}(\theta)-y)\hat{W}_y\hat{\epsilon}(1+\sqrt{n}|\theta-\theta_y|)\\
(6) = & \quad -\tfrac{1}{2}(\hat{g}(\theta)-y)'\left(\hat{W}(\theta)-\hat{W}_y\right)(\hat{g}(\theta)-y) + \tfrac{1}{2}(g(\theta)-y)'(W(\theta)-W_y)(g(\theta)-y).
\end{aligned}$$

We will bound each of these terms (in order of magnitude), so that each term is either $o_P\left(n^{-1}\right)$ or satisfies condition (60).

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

Using Assumption 2.5,

$$\frac{\sqrt{n}|(1.2)|}{|\theta-\theta_y|(1+\sqrt{n}|\theta-\theta_y|)} \leq \sup_{\theta\in N(\theta_0),y\in\mathcal{Y}} |\hat{W}(\theta)-W_y| \frac{|g(\theta)-g(\theta_y)|^2}{|\theta-\theta_y|^2} = o_P(1).$$

Next note that (1.1) will cancel later, where

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

The second term

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

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

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

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

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

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

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

which cancels (1.1), and

$$(3.2.2) = -\left(g\left(\theta_y\right) - y\right)' \left(\hat{W}_y - W_y\right)\left(g\left(\theta\right) - g\left(\theta_y\right)\right)$$

Under Assumption 3, $g\left(\theta_y\right) - y \equiv 0$ so (3.2.2) disappears. Under both Assumption 4 and 5,

$$\sup_{y \in \mathcal{Y}} \sqrt{n}|\hat{W}_y - W_y| = O_P\left(1\right). \tag{62}$$

Since there is also $|g\left(\theta_y\right) - y| = o\left(1\right)$, we conclude that

$$\frac{\sqrt{n}|\left(3.2.2\right)|}{|\theta - \theta_y|\left(1 + \sqrt{n}|\theta - \theta_y|\right)} \leq \sqrt{n}|\hat{W}_y - W_y| \frac{|g\left(\theta\right) - g\left(\theta_y\right)|}{|\theta - \theta_y|} |g\left(\theta_y\right) - y| = o_P\left(1\right). \tag{63}$$

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

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

and

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

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

We will next deal with (6) first before dealing with (5). First under Assumption 5, $(6) \equiv 0$ since $\hat{W}\left(\theta\right) = \hat{W}_y = \hat{W}$, and $W\left(\theta\right) = W_y = W$. Next, under Assumption 4, (6) is to the first order approximately

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

Finally, under Assumption 3, $g(\theta_y) = y$, then we write the second term in (6) as

$$(g(\theta) - g(\theta_y))' (W(\theta) - W_y)(g(\theta) - g(\theta_y)) = o\left(|\theta - \theta_y|^2\right).$$

Also write the first term in (6) as

$$\begin{aligned} & (\hat{g}(\theta) - g(\theta_y))' \left(\hat{W}(\theta) - \hat{W}_y\right)(\hat{g}(\theta) - g(\theta_y)) \\ = & \left(\hat{g}(\theta) - \hat{g}(\theta_y) + O_P\left(\tfrac{1}{\sqrt{n}}\right)\right)' \left(\hat{W}(\theta) - \hat{W}_y\right)\left(\hat{g}(\theta) - \hat{g}(\theta_y) + O_P\left(\tfrac{1}{\sqrt{n}}\right)\right) \end{aligned}$$

Futhermore, note that

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

Since $\hat{\epsilon} = o_P\left(\frac{1}{\sqrt{n}}\right)$, the first part of (6) (and thus the entire (6)) satisfies

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

Finally, consider

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

$$\begin{aligned} (5.1) &= (\hat{g}(\theta) - g(\theta_y))\hat{W}_y\hat{\epsilon}\left(1 + \sqrt{n}|\theta - \theta_y|\right) \\ &= \left(O_P\left(\frac{1}{\sqrt{n}}\right) + O(|\theta - \theta_y|)\right) o_P\left(\frac{1}{\sqrt{n}}\right)\left(1 + \sqrt{n}|\theta - \theta_y|\right) = o_P\left(n^{-1}\right) + o_P\left(\frac{1}{\sqrt{n}}|\theta - \theta_y|\right) + o_P\left(|\theta - \theta_y|^2\right). \end{aligned}$$

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. Under Assumption 3, $(5.2) \equiv 0$ since $g(\theta_y) = y$. Under Assumption 4,

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

then we can write, as required,

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

Finally, under Assumption 5, where $\sup_{y\in\mathcal{Y}}|y| = o\left(n^{-\frac{1}{4}}\right)$, so that

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

Then we can write, by Cauchy-Schwartz,

$$(5.2) = o\left(n^{-1/4}\right)\hat{W}_y O_P\left(\frac{\sqrt{|\theta-\theta_y|}}{\sqrt{n}}\right) = o_P\left(\frac{1}{\sqrt{n}}\times\frac{\sqrt{|\theta-\theta_y|}}{n^{1/4}}\right) = o_P\left(\frac{1}{n}+\frac{|\theta-\theta_y|}{\sqrt{n}}\right).$$

By now we have fully verified (58). ■

Next we describe locally uniform versions of the convergence results in Chernozhukov and Hong (2003). Let $h_y = \sqrt{n}\left(\theta - \check{\theta}_y\right)$, and let

$$p_n^y(h_y) = \frac{1}{\sqrt{n}^k}\frac{\pi_2\left(\check{\theta}_y + \frac{h_y}{\sqrt{n}}\right)\exp\left(n\hat{Q}_y\left(\check{\theta}_y + \frac{h_y}{\sqrt{n}}\right)\right)}{\int \pi_2(\theta)\exp\left(n\hat{Q}_y(\theta)\right)d\theta} \quad \text{where} \quad \pi_2(\theta) = \pi(\theta)\det\left(\hat{\Sigma}(\theta)\right)^{-1/2} \tag{64}$$

and let $p_\infty^y(h_y) = \frac{\sqrt{|J_y|}}{\sqrt{(2\pi)^k}}\cdot e^{-\frac{1}{2}h_y' J_y h_y}$, as well as $||f||_{TMV(\alpha)} = \int(1+|h|^\alpha)|f(h)|dh$. Also let $\bar{\eta}^y$ and $\bar{\eta}_\tau^y$ be defined through

$$\bar{\eta}^y = \frac{\int \rho(\theta)\pi_2(\theta)e^{n\hat{Q}_y(\theta)}d\theta}{\int \pi_2(\theta)e^{n\hat{Q}_y(\theta)}d\theta}, \quad \text{and} \quad \int_{1\left(\theta:\rho(\theta)\le\bar{\eta}_\tau^y\right)}\pi_2(\theta)e^{n\hat{Q}_y(\theta)}d\theta = \tau\int\pi_2(\theta)e^{n\hat{Q}_y(\theta)}d\theta.$$

LEMMA 5 Let the conditions in Lemma 2 and (58) and (59) in Lemma 3 hold, then for any $0 \le \alpha \le \infty$,

$$\sup_{y\in\mathcal{Y}}||p_n^y(h_y) - p_\infty^y(h_y)||_{TMV(\alpha)} = o_P(1). \tag{65}$$

If $\eta(\theta)$ is twice continuously and boundedly differentiable, then

$$\sup_{y\in\mathcal{Y}}\left|\sqrt{n}\left(\bar{\eta}_y - \eta\left(\check{\theta}_y\right)\right)\right| = o_P(1). \tag{66}$$

For any $\tau\in(0,1)$, and q_τ being the τth percentile of $N(0,1)$,

$$\sup_{y\in\mathcal{Y}}\left|\bar{\eta}_\tau^y - \eta\left(\check{\theta}_y\right) - q_\tau\frac{1}{\sqrt{n}}\sqrt{\Delta_\theta\eta(\theta_y)' J_y^{-1}\Delta_\theta\eta(\theta_y)}\right| = o_P\left(\frac{1}{\sqrt{n}}\right), \tag{67}$$

If the information matrix equality holds, then (5) holds. Furthermore, under Assumption 2,

$$\sup_{y\in\mathcal{Y}}\left|\sqrt{n}\left(\bar{\eta}^y - \bar{\eta}^0\right) - \sqrt{n}\left(\eta_y - \eta_0\right)\right| = o_P(1). \quad \text{where} \quad \eta_y = \eta(\theta_y). \tag{68}$$

Likewise when $\bar{\eta}^y$ and $\bar{\eta}^0$ are replaced by $\bar{\eta}_\tau^y$ and $\bar{\eta}_\tau^0$:

$$\sup_{y\in\mathcal{Y}}\left|\sqrt{n}\left(\bar{\eta}_\tau^y - \bar{\eta}_\tau^0\right) - \sqrt{n}\left(\eta_y - \eta_0\right)\right| = o_P(1). \quad \text{where} \quad \eta_y = \eta(\theta_y). \tag{69}$$

Proof of Lemma 5 : First write

$$p_n^y(h) = \frac{\pi_2\left(\frac{h}{\sqrt{n}} + \check{\theta}_y\right)\exp\left(\omega_y(h)\right)}{\int_{H_n}\pi_2\left(\frac{h}{\sqrt{n}} + \check{\theta}_y\right)\exp\left(\omega_y(h)\right)dh} = \frac{\pi_2\left(\frac{h}{\sqrt{n}} + T_n\right)\exp\left(\omega_y(h)\right)}{C_n^y},$$

where

$$\omega_y(h) = n\left(\hat{Q}_y\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) - \hat{Q}_y\left(\check{\theta}_y\right)\right)$$

and

$$C_n^y \equiv \int \pi_2\left(\frac{h}{\sqrt{n}} + \check{\theta}_y\right)\exp\left(\omega_y(h)\right)dh. \tag{70}$$

We will show that for each $\alpha \geq 0$,

$$A_{1n} \equiv \sup_{y\in\mathcal{Y}}\int |h|^\alpha \left|\exp\left(w_y(h)\right)\pi_2\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) - \exp\left(-\frac{1}{2}h'J_yh\right)\pi_2\left(\theta_y\right)\right|dh \xrightarrow{p} 0. \tag{71}$$

Given (71), taking $\alpha = 0$ we have

$$\sup_{y\in\mathcal{Y}}\left|C_n^y - \int_{\mathbb{R}^k} e^{-\frac{1}{2}h'J_yh}\pi_2\left(\theta_y\right)dh\right| = \sup_{y\in\mathcal{Y}}\left|C_n^y - \pi_2\left(\theta_y\right)(2\pi)^{\frac{k}{2}}\det|J_y|^{-1/2}\right| = o_P(1). \tag{72}$$

Next note that

$$\int |h|^\alpha\left|p_n^y(h) - p_\infty^y(h)\right| = A_n^y \cdot C_{n,y}^{-1},$$

where

$$A_n^y \equiv \int |h|^\alpha\left|e^{w_y(h)}\pi_2\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) - (2\pi)^{-k/2}|J_y|^{1/2}\exp\left(-\frac{1}{2}h'J_yh\right)\cdot C_n^y\right|dh.$$

Using (72), to show (65) it suffices to show that uniformly in $y \in \mathcal{Y}$, $A_n^y \xrightarrow{p} 0$. But

$$A_n^y \leq A_{1n}^y + A_{2n}^y$$

where by (71) $\sup_{y\in\mathcal{Y}} A_{1n}^y \xrightarrow{p} 0$, and by (72), uniformly in $y \in \mathcal{Y}$,

$$\begin{aligned}
A_{2n}^y &= \int |h|^\alpha\left|C_n^y(2\pi)^{-k/2}|J_y|^{1/2}\exp\left(-\frac{1}{2}h'J_yh\right) - \pi_2\left(\theta_y\right)\exp\left(-\frac{1}{2}h'J_yh\right)\right|dh \\
&= \left|C_n^y(2\pi)^{-k/2}|J_y|^{1/2} - \pi_2\left(\theta_y\right)\right|\int |h|^\alpha\exp\left(-\frac{1}{2}h'J_yh\right)dh \xrightarrow{p} 0.
\end{aligned}$$

We now show (71). Use (58) and the conclusion of Lemma 3 to write, for any $\delta \to 0$ and $|h| \leq \sqrt{n}\delta$,

$$\begin{aligned}
\omega_y(h) =& n\left(\check{\theta}_y - \theta_y + \frac{h}{\sqrt{n}}\right)\Delta_n^y - \frac{1}{2}\left(\check{\theta}_y - \theta_y + \frac{h}{\sqrt{n}}\right)'J_y\left(\check{\theta}_y - \theta_y + \frac{h}{\sqrt{n}}\right) + R_n^y\left(\theta_y, \check{\theta}^y + \frac{h}{\sqrt{n}}\right) \\
&- n\left(\left(\check{\theta}_y - \theta_y\right)\Delta_n^y - \frac{1}{2}\left(\check{\theta}_y - \theta_y\right)'J_y\left(\check{\theta}_y - \theta_y\right) + R_n^y\left(\theta_y, \check{\theta}^y\right)\right) \\
=& h'\sqrt{n}\Delta_n^y - \frac{1}{2}h'J_yh - \sqrt{n}\left(\check{\theta}_y - \theta_y\right)'J_yh + nR_n^y\left(\theta_y, \check{\theta}^y + \frac{h}{\sqrt{n}}\right) + nR_n^y\left(\theta_y, \check{\theta}^y\right) \\
=& -\frac{1}{2}h'J_yh + o_P(1)h + o_P(1)\left(h + \sqrt{n}\left(\check{\theta}_y - \theta_y\right)\right)^2 + o_P(1)\left|h + \sqrt{n}\left(\check{\theta}_y - \theta_y\right)\right| \\
=& -\frac{1}{2}h'J_yh + o_P(1)h + o_P(1)h^2 + o_P(1).
\end{aligned}$$

Then we can bound (71) $\leq B_{1n}^y + B_{2n}^y$, where

$$B_{n1}^y = \int_{|h|\leq\sqrt{n}\delta} |h|^\alpha e^{-\frac{1}{2}h'J_y h} \left| e^{o_P(1)h+o_P(1)h^2+o_P(1)} \pi_2\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) - \pi_2\left(\theta_y\right) \right| dh$$

and

$$B_{n2}^y = \int_{|h|\geq\sqrt{n}\delta} \int |h|^\alpha \left| \exp\left(w_y\left(h\right)\right) \pi_2\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) - \exp\left(-\frac{1}{2}h'J_y h\right) \pi_2\left(\theta_y\right) \right| dh.$$

Further bound $B_{1n}^y \leq B_{11n}^y + B_{21n}^y + B_{31n}^y$, where

$$B_{11n}^y = \int |h|^\alpha e^{-\frac{1}{2}h'J_y h} dh \sup_{|h|\leq M} \left| e^{o_P(1)h+o_P(1)h^2+o_P(1)} \pi_2\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) - \pi_2\left(\theta_y\right) \right|,$$

$$B_{21n}^y = \int_{|h|\geq M} |h|^\alpha e^{-\frac{1}{2}h'J_y h} e^{o_P(1)h+o_P(1)h^2+o_P(1)} \pi_2\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) dh$$

$$B_{31n}^y = \pi_2\left(\theta_y\right) \int_{|h|\geq M} |h|^\alpha e^{-\frac{1}{2}h'J_y h} dh.$$

Since $\sup_{y\in\mathcal{Y}} B_{11n}^y = o_P\left(1\right)$ holds for each fixed M, it also holds for some sequence of $M \to \infty$. But for any $M \to \infty$, both $\sup_{y\in\mathcal{Y}} B_{21n}^y = o_P\left(1\right)$ and $\sup_{y\in\mathcal{Y}} B_{31n}^y = o_P\left(1\right)$ since $e^{-\frac{1}{2}h'J_y h}$ eventually dominates. Therefore $\sup_{y\in\mathcal{Y}} B_{1n}^y = o_P\left(1\right)$. Next we bound $B_{2n}^y \leq B_{12n}^y + B_{22n}^y$, where for each $\delta > 0$ (and hence for some sequence of $\delta \to 0$) as $n \to \infty$:

$$B_{12n}^y = \int_{|h|\geq\sqrt{n}\delta} \int |h|^\alpha \exp\left(-\frac{1}{2}h'J_y h\right) \pi_2\left(\theta_y\right) dh, \qquad \sup_{y\in\mathcal{Y}} B_{12n}^y = o\left(1\right).$$

Change variable to $\theta = \check{\theta} + h/\sqrt{n}$ and recall $\omega_y\left(h\right)$, write

$$\begin{aligned} B_{22n}^y =& \int_{|h|\geq\sqrt{n}\delta} \int |h|^\alpha \exp\left(w_y\left(h\right)\right) \pi_2\left(\check{\theta}_y + \frac{h}{\sqrt{n}}\right) dh \\ =& \sqrt{n}^{k+\alpha} \int_{|\theta-\check{\theta}_y|\geq\delta} |\theta - \check{\theta}_y|^\alpha \exp\left(n\left(\hat{Q}_y\left(\theta\right) - \hat{Q}_y\left(\check{\theta}_y\right)\right)\right) \pi_2\left(\theta\right) d\theta. \end{aligned}$$

It is easy to see that under the conditions in Lemma 2, $\exists \epsilon > 0$ (given δ),

$$P\left(\sup_{y\in\mathcal{Y}} \sup_{|\theta-\theta_y|\geq\delta} \left(\hat{Q}_y\left(\theta\right) - \hat{Q}_y\left(\check{\theta}_y\right)\right) \leq -\epsilon\right) \to 1.$$

On this event,

$$B_{22n}^y \leq C\sqrt{n}^{\alpha+k} e^{-n\epsilon} \int |\theta - \check{\theta}_y|^\alpha \pi_2\left(\theta\right) d\theta = o_P\left(1\right).$$

This completes the proof for (65).

Next consider (66). Write $\bar{\eta}_y = \int \rho(\theta)\, p_n^y(\theta)\, d\theta = \int \rho\left(\check{\theta}_y + h/\sqrt{n}\right) p_n^y(h)\, dh$. Therefore

$$(66) = \int \sqrt{n}\left(\rho\left(\check{\theta}_y + h/\sqrt{n}\right) - \rho\left(\check{\theta}_y\right)\right) p_n^y(h)\, dh = (1)_y + (2)_y$$

where $(1)_y = \frac{\partial}{\partial\theta}\rho\left(\check{\theta}_y\right)\int h\left(p_n^y(h) - p_\infty^y(h)\right) dh$ and $(2)_y = \frac{1}{\sqrt{n}}\int h'\rho^{(2)}\left(\check{\theta}_y, h/\sqrt{n}\right) h p_n^y(h)\, dh$, with

$$\sup_{y\in\mathcal{Y}} |(1)_y| \leq \sup_{y\in\mathcal{Y}} \left|\frac{\partial}{\partial\theta}\rho\left(\check{\theta}_y\right)\right| \sup_{y\in\mathcal{Y}} \left|\int h\left|\left(p_n^y(h) - p_\infty^y(h)\right)\right| dh = o_P(1)$$

because of (65). Next $(2)_y$ can be bounded by, for some large $M < \infty$, again using (65)

$$\sup_{y\in\mathcal{Y}} |(2)_y| \leq \frac{M}{\sqrt{n}}\int |h|^2 p_n^y(h)\, dh = \frac{1}{\sqrt{n}} O_P(1) = o_P(1).$$

Define

$$P_y\left(\sqrt{n}\left(\eta(\theta) - \eta\left(\check{\theta}_y\right)\right) \leq s|\mathcal{X}_n\right) \equiv \int_{\eta(\theta)\leq\eta\left(\check{\theta}_y\right)+\frac{s}{\sqrt{n}}} p_n^y(\theta)\, d\theta = \int_{\eta\left(\check{\theta}_y+\frac{h}{\sqrt{n}}\right)\leq\eta\left(\check{\theta}_y\right)+\frac{s}{\sqrt{n}}} p_n^y(h)\, dh$$

We will show the following conditional Delta method, for any compact S,

$$\sup_{y\in\mathcal{Y}}\sup_{s\in S}\left|P_y\left(\sqrt{n}\left(\eta(\theta) - \eta\left(\check{\theta}_y\right)\right) \leq s|\mathcal{X}_n\right) - \int_{\frac{\partial}{\partial\theta}\eta(\theta_y)'h\leq s} p_\infty^y(h)\, dh\right| = o_P(1), \tag{73}$$

where $\int_{\frac{\partial}{\partial\theta}\eta(\theta_y)'h\leq s} p_\infty^y(h)\, dh = \Phi\left(\frac{s}{\sqrt{\frac{\partial}{\partial\theta}\eta(\theta_y)' J_y^{-1}\frac{\partial}{\partial\theta}\eta(\theta_y)}}\right)$. First, immediately from (65)

$$\sup_{y\in\mathcal{Y}}\sup_{s\in S}\left|P_y\left(\sqrt{n}\left(\eta(\theta) - \eta\left(\check{\theta}_y\right)\right) \leq s|\mathcal{X}_n\right) - \int_{\eta\left(\check{\theta}_y+\frac{h}{\sqrt{n}}\right)\leq\eta\left(\check{\theta}_y\right)+\frac{s}{\sqrt{n}}} p_\infty^y(h)\, dh\right| = o_P(1).$$

For $Z \sim N(0, I_k)$, and $X_y = J_y^{-1/2} Z$, we can write, for mean values $\theta\left(\check{\theta}_y, X_y\right)$,

$$\begin{aligned}
&\int_{\eta\left(\check{\theta}_y+\frac{h}{\sqrt{n}}\right)\leq\eta\left(\check{\theta}_y\right)+\frac{s}{\sqrt{n}}} p_\infty^y(h)\, dh = P\left(\sqrt{n}\left(\eta\left(\check{\theta}_y + X_y/\sqrt{n}\right) - \eta\left(\check{\theta}_y\right)\right) \leq s|\check{\theta}_y\right) \\
= \quad & P\left(\eta^{(1)}\left(\check{\theta}_y\right)' X_y \leq s - \tfrac{1}{\sqrt{n}} X_y' \eta^{(2)}\left(\theta\left(\check{\theta}_y, X_y\right)\right) X_y \check{\theta}_y\right) \\
= \quad & P\left(\eta^{(1)}(\theta_y)' X_y \leq s - \tfrac{1}{\sqrt{n}} X_y' \eta^{(2)}\left(\theta\left(\check{\theta}_y, X_y\right)\right) X_y - \left(\eta^{(1)}\left(\check{\theta}_y\right) - \eta^{(1)}(\theta_y)\right)' X_y|\check{\theta}_y\right) \\
= \quad & P\left(\eta^{(1)}(\theta_y)' J_y^{-1/2} Z \leq s - \tfrac{1}{\sqrt{n}} Z' J_y^{-1/2} \eta^{(2)}\left(\theta\left(\check{\theta}_y, X_y\right)\right) J_y^{-1/2} Z - \left(\eta^{(1)}\left(\check{\theta}_y\right) - \eta^{(1)}(\theta_y)\right)' J_y^{-1/2} Z|\check{\theta}_y\right)
\end{aligned}$$

Since $\eta(\cdot)$ has bounded 2nd derivative and $\forall\epsilon > 0$, $\exists M < \infty$ such that $P(|Z| > M) < \epsilon$, and that $sup_{y\in\mathcal{Y}}|\eta^{(1)}\left(\check{\theta}_y\right) - \eta^{(1)}(\theta_y)| = o_P(1)$, for some $C > 0$, we can write

$$\begin{aligned}
&\left|\int_{\eta\left(\check{\theta}_y+\frac{h}{\sqrt{n}}\right)\leq\eta\left(\check{\theta}_y\right)+\frac{s}{\sqrt{n}}} p_\infty^y(h)\, dh - P\left(\eta^{(1)}(\theta_y)' J_y^{-1/2} Z \leq s\right)\right| \\
&\leq 2P(|Z| > M) + 2P\left(s - \tfrac{CM^2}{\sqrt{n}} - o_P(1) M \leq \eta^{(1)}(\theta_y)' J_y^{-1/2} Z \leq s + \tfrac{CM^2}{\sqrt{n}} + o_P(1) M\right)
\end{aligned}$$

For any given $0 < M < \infty$, it follows from $\inf_{y\in\mathcal{Y}} \eta^{(1)}(\theta_y)' J_y^{-1}\eta^{(1)}(\theta_y) > 0$, and hence $\eta^{(1)}(\theta_y)' J_y^{-1/2} Z$ having uniformly bounded density, that

$$\sup_{y\in\mathcal{Y}}\sup_{s\in S} P\left(s - \frac{CM^2}{\sqrt{n}} - o_P(1)M \leq \eta^{(1)}(\theta_y)' J_y^{-1/2} Z \leq s + \frac{CM^2}{\sqrt{n}} + o_P(1)M\right) = o_P(1).$$

Hence we have proven (73), which we now use to show (67) using relatively standard arguments. The goal is to convert, for now $T_y \equiv \eta^{(1)}(\theta_y)' J_y^{-1/2} Z$,

$$\sup_{y\in\mathcal{Y}}\sup_{s\in S}\left|P(h_y \leq s|\mathcal{X}_n) - P(T_y \leq s)\right| = o_P(1) \tag{74}$$

where $h_y \sim p_n^y(h_y)$ into a re-expression of (67), for all $\epsilon > 0$,

$$P\left(\sup_{y\in\mathcal{Y}}\left|Q_\tau(h_y) - Q_\tau(T_y)\right| > \epsilon\right) = o(1). \tag{75}$$

To simplify notation let $\hat{F}_y(s) = P(h_y \leq s|\mathcal{X}_n)$, $F_y(s) = P(T_y \leq s)$, $\hat{Q}_y(\tau) = Q_\tau(h_y)$, $Q_y(\tau) = Q_\tau(T_y)$, then note that by uniform (in y) strict monotonicity of $F_y(s)$ in s, $\exists\delta > 0$ such that

$$\sup_{y\in\mathcal{Y}} F_y(Q_y(\tau) - \epsilon) \leq \tau - \delta, \quad \inf_{y\in\mathcal{Y}} F_y(Q_y(\tau) + \epsilon) \geq \tau + \delta$$

Furthermore $|\hat{Q}_y(\tau) - Q_y(\tau)| > \epsilon$ implies either

$$\hat{F}_y(Q(\tau) - \epsilon) \geq \tau \Longrightarrow \hat{F}_y(Q(\tau) - \epsilon) - F_y(Q_y(\tau) - \epsilon) \geq \delta$$

or $\hat{F}_y(Q(\tau) + \epsilon) \leq \tau \Longrightarrow \hat{F}_y(Q(\tau) + \epsilon) - F_y(Q_y(\tau) + \epsilon) \leq -\delta$. Therefore

$$P\left(\sup_{y\in\mathcal{Y}}|\hat{Q}_y(\tau) - Q_y(\tau)| > \epsilon\right) \leq P\left(\sup_{y\in\mathcal{Y}}\sup_{s\in S}|\hat{F}_y(s) - F_y(s)| > \delta\right) \to 0.$$

Now (67) is proven. Finally we now show (68). First applying the Delta method to the conclusion of Lemma 3 we have

$$\sup_{y\in\mathcal{Y}}|\sqrt{n}\left(\eta\left(\check{\theta}_y\right) - \eta(\theta_y)\right) - \Delta_\theta\eta(\theta_y)' J_y^{-1}\sqrt{n}\Delta_n^y| = o_P(1).$$

Next use this and (66) to write (for $o_P(1)$ uniform in $y \in \mathcal{Y}$),

$$\sqrt{n}\left(\bar{\eta}^y - \bar{\eta}^0\right) - \sqrt{n}(\eta(\theta_y) - \eta_0) = \sqrt{n}\left(\eta\left(\check{\theta}_y\right) - \eta(\theta_y) - \left(\eta\left(\check{\theta}_0\right) - \eta_0\right)\right) + o_P(1).$$

To show that

$$\sqrt{n}\left(\eta\left(\check{\theta}_y\right) - \eta(\theta_y) - \left(\eta\left(\check{\theta}_0\right) - \eta_0\right)\right) = o_P(1), \tag{76}$$

write it as

$$\begin{aligned}
&\Delta_\theta \eta(\theta_y)' J_y^{-1} \sqrt{n}\Delta_n^y - \Delta_\theta \eta(\theta_0)' J_0^{-1} \sqrt{n}\Delta_n^0 + o_P(1) \\
=& \left(\Delta_\theta \eta(\theta_y)' J_y^{-1} - \Delta_\theta \eta(\theta_0)' J_0^{-1}\right) \sqrt{n}\Delta_n^0 + \Delta_\theta \eta(\theta_0)' J_0^{-1} \left(\sqrt{n}\Delta_n^y - \sqrt{n}\Delta_n^0\right) \\
&+ \left(\Delta_\theta \eta(\theta_y)' J_y^{-1} - \Delta_\theta \eta(\theta_0)' J_0^{-1}\right) \left(\sqrt{n}\Delta_n^y - \sqrt{n}\Delta_n^0\right) + o_P(1).
\end{aligned}$$

Since $\sup_{y\in\mathcal{Y}} |\Delta_\theta \eta(\theta_y)' J_y^{-1} - \Delta_\theta \eta(\theta_0)' J_0^{-1}| = o_P(1)$, it suffices to show $\sup_{y\in\mathcal{Y}} |\sqrt{n}\Delta_n^y - \sqrt{n}\Delta_n^0| = o_P(1)$. Under Assumption 2, $\Delta_n^y = (\hat{g}(\theta_y) - g(\theta_y))' W_y G_y$ as in the proof of Lemma 4, so that by the same arguments, we only need $\sup_{y\in\mathcal{Y}} |\sqrt{n}(\hat{g}(\theta_y) - g(\theta_y) - \hat{g}(\theta_0) + g(\theta_0))| = o_P(1)$, which is Assumption 2. Same arguments above apply to replace $\bar{\eta}_y$ by $\bar{\eta}_\tau^y$ by using (67) instead of (66).

Likewise, Assumption 2 also implies that

$$\sup_{y\in\mathcal{Y}} |\Delta_\theta \eta(\theta_y)' J_y^{-1} \sqrt{n}\Delta_n^y - \Delta_\theta \eta(\theta_0)' J_0^{-1} \sqrt{n}\Delta_n^0| = o_P(1). \tag{77}$$

Next we combine (77), (67) and the conclusion of Lemma (3) to write that

$$\sup_{y\in\mathcal{Y}} |\sqrt{n}(\bar{\eta}_\tau^y - \eta^y) - \Delta_\theta \eta(\theta_0)' J_0^{-1} \sqrt{n}\Delta_n^0 - q_\tau \sqrt{\Delta_\theta \eta(\theta_y)' J_y^{-1} \Delta_\theta \eta(\theta_y)}| = o_P(1).$$

Then the posterior coverage validity in (5) follows from

$$\begin{aligned}
&\sup_{y\in\mathcal{Y}} |P\left(\sqrt{n}(\bar{\eta}_\tau^y - \eta^y) \leq 0\right) - (1-\tau)| \\
=&\sup_{y\in\mathcal{Y}} |P\left(\Delta_\theta \eta(\theta_0)' J_0^{-1} \sqrt{n}\Delta_n^0 + q_\tau \sqrt{\Delta_\theta \eta(\theta_y)' J_y^{-1} \Delta_\theta \eta(\theta_y)} + o_P(1) \leq 0\right) - (1-\tau)| = o_P(1),
\end{aligned}$$

since $\Delta_\theta \eta(\theta_0)' J_0^{-1} \sqrt{n}\Delta_n^0 + o_P(1) \rightsquigarrow N\left(0, \Delta_\theta \eta(\theta_0)' J_0^{-1} \Delta_\theta \eta(\theta_0)\right)$ and $\sup_{y\in\mathcal{Y}} |\Delta_\theta \eta(\theta_y)' J_y^{-1} \Delta_\theta \eta(\theta_y) - \Delta_\theta \eta(\theta_0)' J_0^{-1} \Delta_\theta \eta(\theta_0)| = o(1)$. ■

B.4 Asymptotic Indirect Inference Likelihood

Creel and Kristensen (2011) demonstrated that the indirect inference likelihood function asymptotes to the continuously updating GMM criterion function. Consider, for $f_{\Gamma_n}(\cdot|\theta)$ denoting the density of T_n given θ,

$$f_n(\theta|T_n + y) = \frac{f_{\Gamma_n}(T_n + y|\theta)\pi(\theta)}{\int_\Theta f_{\Gamma_n}(T_n + y|\theta)\pi(\theta)\, d\theta} = \frac{e^{n\hat{Q}_1^y(\theta)}\pi_2(\theta)}{\int_\Theta e^{n\hat{Q}_1^y(\theta)}\pi_2(\theta)\, d\theta},$$

where we define $\pi_2(\theta) = \pi(\theta) \det(\Sigma(\theta))^{-1/2}$ and

$$\hat{Q}_1^y(\theta) = \frac{1}{n} \log f_{\Gamma_n}(T_n + y|\theta) - \frac{d}{2}\frac{\log n}{n} + \frac{d \log\sqrt{2\pi}}{n} + \frac{1}{2n} \log\det(\Sigma(\theta))$$

Also let $\hat{Q}_2^y(\theta) = -\frac{1}{2}(T_n + y - t(\theta))' \Sigma(\theta)^{-1}(T_n + y - t(\theta))$. We will show the following two conditions:

$$\sup_{\theta\in\Theta}\sup_{y\in\mathcal{Y}} |\hat{Q}_1^y(\theta) - \hat{Q}_2^y(\theta)| = o_P(1), \tag{78}$$

and for any $\delta \to 0$,

$$\sup_{y\in\mathcal{Y}}\sup_{|\theta-\theta_y|\leq\delta} \frac{n|\hat{Q}_1^y(\theta) - \hat{Q}_2^y(\theta)|}{1+\sqrt{n}|\theta-\theta_y| + n|\theta-\theta_y|^2} = o_P(1). \tag{79}$$

Since Assumptions 2 and one of 3 or 4, hence the conditions in Lemmas 2 and 3, hold for $\hat{Q}_2^y(\theta)$, because of (78) and (79) they also hold for $\hat{Q}_1^y(\theta)$. Define

$$f_{Z_n}(z|\theta) = \sqrt{n}^{-d} \det(\Sigma(\theta))^{1/2} f_{\Gamma_n}\left(\frac{1}{\sqrt{n}}\Sigma(\theta)^{/1/2} z + t(\theta)\,|\theta\right)$$

In other words, $f_{Z_n}(z|\theta)$ is the density of $Z_n = \sqrt{n}\Sigma(\theta)^{-1/2}(T_n - t(\theta))$ at $Z_n = z$. The following lemma formalizes a notion that $Z_n \xrightarrow{d} N(0, I)$ uniformly in θ and mirrors Assumption 1 in Creel and Kristensen (2011) who also provided verification in some examples.

LEMMA 6 (78) and (79) both hold if the following two conditions hold:

1. $\sup_{\theta\in\Theta}\sup_{z\in R^d} |\log f_{Z_n}(z|\theta) - \log\frac{1}{\sqrt{2\pi}^d} e^{-\frac{1}{2}z'z}| = o(n)$.
2. For any $\delta_1 \to 0$, $\delta_2 \to 0$,

$$\sup_{|\theta-\theta_0|\leq\delta_1}\sup_{|z|\leq\sqrt{n}\delta_2} \frac{|\log f_{Z_n}(z|\theta) - \log\frac{1}{\sqrt{2\pi}^d} e^{-\frac{1}{2}z'z}|}{1+|z|+|z|^2} = o(1)$$

Proof: Write $f_{\Gamma_n}(x|\theta) = \sqrt{n}^d \det(\Sigma(\theta))^{-1/2} f_{Z_n}\left(\sqrt{n}\Sigma(\theta)^{-1/2}(x - t(\theta))\,|\theta\right)$. Therefore

$$\hat{Q}_1^y(\theta) = \frac{d\log\sqrt{2\pi}}{n} + \frac{1}{n}\log f_{Z_n}\left(\sqrt{n}\Sigma(\theta)^{-1/2}(T_n + y - t(\theta))\,|\theta\right)$$

Then (78) is an immediate consequence of the first condition. Next we use the second condition to show (79). Since $T_n \xrightarrow{p} t(\theta_0)$, for any $\delta_1 \to 0$, find $\delta_2 \to 0$ sufficiently slowly such that $w.p. \to 1$,

$$\sup_{y\in\mathcal{Y}}\sup_{|\theta-\theta_y|\leq\delta_1} |\Sigma(\theta)^{-1/2}(T_n + y - t(\theta))| \leq \delta_2$$

Hence we can invoke condition 2 on this event sequence, and also use $T_n - t(\theta_0) = O_P\left(\frac{1}{\sqrt{n}}\right)$, to bound $n\left(\hat{Q}_1^y(\theta) - \hat{Q}_2^y(\theta)\right)$ uniformly in $y \in \mathcal{Y}$ and $|\theta - \theta_y| \leq \delta_1$ by

$$o_P\left(1 + \sqrt{n}|t(\theta_0) + y - t(\theta)| + n|t(\theta_0) + y - t(\theta)|^2\right).$$

Finally we note that $|t(\theta_0)+y-t(\theta)| = O(|\theta_y-\theta_0|)$. For example, in an exactly identified model, $t(\theta_0)+y=t(\theta_y)$.

Alternatively, we can also strengthen the second condition to

$$\sup_{|\theta-\theta_0|\leq\delta_1}\sup_{|z|\leq\sqrt{n}\delta_2}|\log f_{Z_n}(z|\theta)-\log\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}z'z}| = o(1).$$

Condition 1 and (78) can also be further relaxed so that $f_{Z_n}(z|\theta)$ is only required to be approximated by the normal density for θ close to θ_0. They can be replaced by the requirement that there exists $Q(y,\theta)$ such that Lemma 2 applies to $\hat{Q}(y,\theta)=\frac{1}{n}\log f_{\Gamma_n}(T_n+y|\theta)$ and $Q(y,\theta)$. □

B.5 Iterative Applications

It is possible to apply BIL and ABC-GMM iteratively in combination with adaptive importance sampling. For example, under assumption 5, which is applicable to overidentified quantile IV methods or simulated method of moments, it can be shown using the same arguments in the proof of Lemma 4 that for any $\sup_{y\in\mathcal{Y}}|y|\to 0$, $|\tilde{\theta}_y-\theta_y| = o_P\left(n^{-\frac{1}{3}}\right)$. An iterative application with at least two steps can possibly reduce computation burden and achieve $\sqrt{n}$ consistency and asymptotic normality.

In the first step, a larger bandwidth $h\to 0$ can be used 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 et al. (2015). For $\sup_{y\in\mathcal{Y}}|y| = o(1)$, the arguments in Theorem 3 in Jun et al. (2015) 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 et al. (2015) 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.

The idea of iteratation dates back to Robinson (1988), who in the context of smooth models with analytic moments showed that a finite number of Gauss-Newton iteration can convert any polynomial rate consistent estimator ($\hat{\theta} - \theta_0 = O_P\left(n^{-\alpha}\right)$ for $0 < \alpha < 1/2$) into $\sqrt{n}$ consistency. The results can also be shown when the Jacobian and Hessian need to be numerically computed under suitable conditions on the step size choice. Obvious our method can also be used as initial inputs to Robinson (1988)'s iteration scheme, or as subsequent iteration steps. If we only desire $n^{-\alpha}$ rate, for $0 < \alpha < 1/2$ in a given step, we would only need $n^{\alpha}h^{p+1} \to 0$ and $Sh^k \to \infty$, implying that $S/\left(n^{\frac{\alpha}{p+1}}\right) \to \infty$.

B.6 A comment on importance sampling

Both BIL and ABC-GMM require the choice of tuning parameters including the kernel function, the bandwidth and the number of simulations. The Metropolis-Hastings MCMC also requires the choice of a proposal density, the step size, and either a number of simulations or an algorithm for monitoring convergence. Other algorithms such as a nonadaptive importance sampler, might only require choosing the number of simulations S. For example, we can define a SL-GMM (simulated Laplace) estimator using (10) as

$$\hat{\theta}_{SL} = \frac{\sum_s \theta^s \pi\left(\theta^s\right) \exp(n\hat{Q}_n(\theta^s))}{\sum_s \pi\left(\theta^s\right) \exp(n\hat{Q}_n(\theta^s))} \quad \text{or when } \pi\left(\theta\right) = c \quad \hat{\theta}_{SL} = \frac{\sum_s \theta^s \exp(n\hat{Q}_n(\theta^s))}{\sum_s \exp(n\hat{Q}_n(\theta^s))} \tag{80}$$

where

$$\hat{Q}_n(\theta^s) = -\frac{1}{2}\hat{g}(\theta^s)' W \hat{g}(\theta^s) \tag{81}$$

When the target density is known, the conventional wisdom of importance sampling includes unbiasedness and a variation of the order of $1/S$. However, the current situation is quite different. First of all, since the target density is not directly known, importance sampling is required to compute both the numerator and the dominator in the ratio that defines the Bayesian posterior mean, and thus has a finite sample bias that will only vanish asymptotically. Second, the variance of the importance sampler can also be larger because of the spiky behavior of the posterior density of the parameters.

Putting the bias issue aside, the following example illustrates the potential difficulty with the importance sampling variance. A full-scale theoretical analysis of importance sampling is beyond the scope of the current paper.

Let $f(\mu) = N\left(\mu_0, \frac{1}{n} I_k\right)$, where $\mu_0 = 2\pi * \ell_k$ for ℓ_k a constant vector of ones but this fact is not known to the importance sampler. The importance sampler draws $\mu_s, s = 1, \ldots, S$ from $\pi(\mu) \sim N(0, I_k)$, and is interested in computing $E\cos(\ell_k'\mu)$ by

$$\hat{\rho} = \frac{1}{S}\sum_{s=1}^{S} \cos\left(\ell_k'\mu_s\right) f(\mu_s)/\pi(\mu_s).$$

Then for any S, as $n \to \infty$, by dominated convergence,

$$E\hat{\rho} = \int \cos\left(2\pi k + \ell_k' z\right) \frac{\sqrt{n}^k}{\sqrt{2\pi}^k} e^{-n\frac{z'z}{2}} dz = \int \cos\left(2\pi k + \ell_k' h/\sqrt{n}\right) \frac{1}{\sqrt{2\pi}^k} e^{-\frac{h'h}{2}} dh \to 1.$$

Next consider the variance however,

$$\begin{aligned} Var(\hat{\rho}) = \tfrac{1}{S}\left(E\cos^2(\ell_k'\mu_s) f^2(\mu_s)/\phi^2(\mu_s) - (E\hat{\rho})^2\right) &= \frac{\int \cos^2\left(\ell_k' u\right) \frac{n^k}{\sqrt{2\pi}^k} e^{-n(\mu-\ell_k)'(\mu-\ell_k)+\frac{1}{2}\mu'\mu} d\mu - (E\hat{\rho})^2}{S} \\ &= \frac{\int \cos^2\left(2\pi k + \ell_k' z\right) \frac{n^k}{\sqrt{2\pi}^k} e^{-nz'z+\frac{1}{2}(z+\ell_k)'(z+\ell_k)} d\mu - (E\hat{\rho})^2}{S} \\ &= \frac{\sqrt{n}^k \int \cos^2\left(2\pi k + \ell_k' h/\sqrt{n}\right) \frac{1}{\sqrt{2\pi}^k} e^{-h'h+\frac{1}{2}(h/\sqrt{n}+\ell_k)'(h/\sqrt{n}+\ell_k)} dh - (E\hat{\rho})^2}{S} \end{aligned}$$

Then by dominated convergence theorem

$$\frac{S}{\sqrt{n}^k} Var(\hat{\rho}) \to \left(\frac{1}{2}\right)^k e^{\frac{1}{2}k}.$$

This suggests that in order for $Var(\hat{\rho}) \to 0$, we would require $\frac{S}{\sqrt{n}^k} \to \infty$, which is a much larger lower bound on S than $S >> n^{\frac{k}{4}}$ or $S >> n^{\frac{k}{2(p+1)}}$. The cost to pay for less tuning parameters is more computation using larger number of simulations S. The general nonlinear case is likely to be more difficult. For example, the denominator $\int \pi(\theta) e^{n\hat{Q}(\theta)} d\theta$ converges to zero at $O_P\left(\frac{1}{\sqrt{n}^k}\right)$ creating numerical instability. If we scale it up by $\frac{1}{\sqrt{n}^k}$ to stabilize the denominator, then its importance sampling variance will explode at the $\sqrt{n}^k$ rate.

In fact we can compare the SL-GMM estimator to a local constant kernel ABC-GMM estimator. Recall that a locally constant ABC-GMM estimator is defined as

$$\hat{\theta}_{LC-ABCGMM} = \frac{\sum_s \theta^s K(y^s/h)}{\sum_s K(y^s/h)}$$

where

$$y^s = \hat{g}(\theta^s) + \frac{1}{\sqrt{n}} W^{-1/2}\xi \tag{82}$$

Suppose that a multivariate normal kernel (ignoring the Jacobian term) is used: $K(z) = \exp(-\frac{1}{2}z'Wz)$, and that the bandwidth is set to $h = 1/\sqrt{n}$. Then

$$K(y^s/h) = \exp(-\frac{n}{2}\left(y^s\right)'Wy^s)$$

Now, using eqn. (82), we can write

$$\left(y^s\right)'Wy^s = \hat{g}(\theta^s)'W\hat{g}(\theta^s) + \frac{2}{\sqrt{n}}\left(\hat{g}(\theta^s)\right)'W^{1/2}\xi + \frac{1}{n}\xi'\xi$$

The first term will dominate, because of the powers of n, so

$$\begin{aligned} K(y^s/h) &\simeq \exp\left(-\frac{n}{2}\hat{g}(\theta^s)'W\hat{g}(\theta^s)\right) \\ &= \exp(n\hat{Q}_n(\theta^s)), \end{aligned}$$

using eqn. 81. Therefore, approximately,

$$\hat{\theta}_{LC-GMM} \simeq \frac{\sum_s \theta^s \exp(n\hat{Q}_n(\theta^s))}{\sum_s \exp(n\hat{Q}_n(\theta^s))},$$

which is the second expression in eqn. 80. Therefore, the SL-GMM estimator is essentially a local constant kernel estimator, with the particularity that a normal kernel is used, and the bandwidth is $h = 1/\sqrt{n}$. In comparison, the ABC-GMM estimator can use a different kernel, a different bandwidth, and local linear or local polynomial nonparametric regression to improve performance.