PLAUSIBLE GMM VIA AVENUE BAYES

[PRELIMINARY VERSION, NOT FOR DISTRIBUTION]

VICTOR CHERNOZHUKOV[a], CHRISTIAN B. HANSEN[b], LINGWEI KONG[c], WEINING WANG[d]

ABSTRACT. Structural estimation in economics often makes use of models formulated in terms of moment conditions. While these moment conditions are generally well-motivated, it is often unknown whether the moment restrictions hold exactly. We consider a quasi-Bayesian approach for performing inference on structural parameters while relaxing the restriction that moment restrictions hold exactly. Within this context, we prove new Bernstein-von Mises (BvM) type theorems for the quasi-posterior distributions, which can be used to obtain tractable approximations in practical settings. We illustrate the approach through simulation and empirical applications. Our applications illustrate that we can obtain informative inference for structural objects, even allowing for substantial relaxations of the requirement that moment conditions hold exactly.

Keywords: sensitivity analysis, misspecification, generalized method of moment(GMM), quasi-Bayes, BvM

1. INTRODUCTION

Moment restrictions are commonly used in the identification and estimation of structural or causal parameters in empirical economics. Prominent examples include instrument exclusion conditions, unconfoundedness assumptions, parallel trend assumptions, and nonlinear moment restrictions imposed in nonlinear structural models. Economists typically use institutional knowledge and economic reasoning to argue for the validity of these restrictions in settings with observational data. Based on these arguments, classical estimation and inference, such as estimation and inference based on the generalized method of moments (GMM), then proceed under the maintained assumption that the posited moment restrictions hold exactly.

While the arguments employed to justify moment restrictions provide a basis for believing that the moment conditions are plausible, they are also generally debatable. That is, it is hard to know that there are no unmodeled sources of confounding or any sources of misspecification such that moment conditions do indeed hold exactly in any given empirical setting. Unfortunately, it is

Date: August 01, 2023.

[a]: Department of Economics, Massachusetts Institute of Technology.
[b]: Booth School of Business, The University of Chicago.
[c]: Department of Economics, Econometrics and Finance, Univeristy of Groningen.
[d]: Department of Economics, Econometrics and Finance, Univeristy of Groningen.

well-known that estimation and inference results obtained under the assumption that moment restrictions hold exactly can be substantially distorted in the sense of returning biased estimates and delivering unreliable conclusions. See, for example, Hansen and Sargent (2001), Hall and Inoue (2003), Hansen and Sargent (2008), Hansen and Sargent (2010), Cheng et al. (2015), and Hansen and Lee (2021).

In this paper, we consider one approach to estimation and inference within a moment condition framework, allowing for the possibility that specified moment conditions do not hold exactly. We consider a setting where a researcher specifies moment conditions $g(\theta) = T^{-1}\sum_t \mathrm{E}[g(Z_t,\theta)]$ for observable data stream $\{Z_t\}$ and parameter of interest θ such that we would have $g(\theta_0) = \mu$ for θ_0 denoting the population value of the parameter of interest. Of course, informative inference about θ_0 is impossible without beliefs about μ. Classical estimation and inference results proceed under the dogmatic prior $\mu \equiv 0$. Rather than adopt dogmatic prior beliefs, we conceptualize the notion that the moment restrictions are plausible – but not known to hold exactly – by assuming the researcher is able to place an informative, but not necessarily dogmatic, prior distribution over μ. The use of a proper prior over μ allows informative inference to proceed while relaxing the usual restriction that $\mu \equiv 0$.

Given that we choose to conceptualize the plausibility of moment conditions by using a proper prior distribution over μ, it is natural to consider estimation and inference based on Quasi-Bayes Laplace type estimators (LTEs) as proposed by Chernozhukov and Hong (2003). Chernozhukov and Hong (2003) argue that LTEs are computationally attractive for moment condition models and verify that they provide accurate frequentist coverage under correctly specified moment conditions.

A technical contribution of our present work is extending the previous results on LTEs to a setting with misspecified moment conditions while allowing for both the number of moments and the number of parameters in the model to increase with the sample size. We provide new large sample approximations for quasi-posterior distributions in this framework. These approximations can be regarded as new Bernstein–von Mises-type theorems for quasi-posterior distributions accounting for additional terms needed to handle potential misspecification. Specifically, we show that the limit distribution for θ conditioning on the misspecification parameter μ follows a Gaussian distribution. The joint distribution is thus a Gaussian mixture. We also verify that the quasi-posterior support concentrates on the frequentist identified region in large samples. This property mimics results for fully Bayesian procedures under partial identification; see, e.g., Gustafson (2010). Finally, we provide an approach to simulating from the Gaussian mixture approximation that may be a useful alternative to simulating directly from the quasi-posterior.

We illustrate our proposed method through simulation exercises and empirical applications. In simulations, we examine performance in practically relevant examples: linear IV, quantile regression, and Difference-in-Difference (DID) models. Unsurprisingly, we find that allowing for potential misspecification by incorporating non-dogmatic priors over μ reflecting potential misspecification produces sets with improved (frequentist) coverage results in the event of model

misspecification. A cost of allowing for potential misspecification by considering non-dogmatic beliefs about μ is that inferential statements must be less precise than those obtained under dogmatic beliefs. We demonstrate via empirical applications that one can still draw economically meaningful conclusions using our approach in real applications under what we believe are sensible beliefs about model misspecification, thus potentially enhancing the credibility of the qualitative empirical results.

We believe our work complements several areas of existing literature. Our formal results contribute to the large literature on posterior concentration results. Many such approximations exist when models are correctly specified; see, for example, Doksum and Lo (1990), Barron et al. (1999), Diaconis and Freedman (1986) and Cox (1993). Shen and Wasserman (2001) compute the rate at which the posterior distribution concentrates around the true parameter value. When there are identification challenges, Moon and Schorfheide (2012) derive a large-sample approximation to the posterior distribution of partially identified structural parameters for models indexed by an identifiable finite dimensional reduced-form parameter vector. See also the review article Gustafson (2010), which provides an overview of results on Bayesian estimation in partially identified settings. Andrews and Mikusheva (2022a) and Andrews and Mikusheva (2022b) examine admissibility and optimality of estimators in the GMM setting under weak identification and verify that quasi-Bayes posterior have desirable properties within this setting.

Our paper also complements a large existing literature on partial identification and sensitivity analysis. See, for example, Bonhomme and Weidner (2022), Armstrong and Kolesár (2021), Masten and Poirier (2020), Chen et al. (2018), Berkowitz et al. (2012), Conley et al. (2012), Chen et al. (2011), Chernozhukov et al. (2007), Imbens and Manski (2004), and Rosenbaum (1987). Within this literature, perhaps the two papers most closely related to our work are Chen et al. (2018) and Armstrong and Kolesár (2021).

Chen et al. (2018) propose Monte Carlo confidence sets for identified sets of parameters using likelihoods and moments in a partially identified model. One important result in Chen et al. (2018) is that quasi-Bayes highest posterior density sets under flat priors over model parameters have correct frequentist coverage in regular but partially identified models. The approach we take is similar to that in Chen et al. (2018), though our assumptions are not entirely encompassed by theirs. For instance, we incorporate the ancillary term μ to help identify θ, and our quasi-posterior distribution is influenced by the specific prior placed upon μ.

Our approach also shares many similarities with Armstrong and Kolesár (2021). Armstrong and Kolesár (2021) employ a frequentist minimax approach to obtaining valid confidence regions for parameters of interest under specified bounds on a deterministic level of misspecification. Their formal results show that their derived confidence regions are near-optimal within a local misspecification framework, and they suggest an approach to performing sensitivity analysis by considering different bounds for the misspecification level. We believe the quasi-Bayesian approach we adopt in this paper complements the fully frequentist GMM approach as in Armstrong and Kolesár (2021). As noted previously, Chernozhukov and Hong (2003) argue that

quasi-Bayesian approaches have desirable computational properties relative to more traditional GMM approaches in some settings. Quasi-Bayes approaches can also be motivated as approximations within fully Bayesian settings as in, e.g., Kim (2002), Gallant (2016), Florens and Simoni (2021), and Andrews and Mikusheva (2022b). Within this context, we believe some researchers may wish to employ proper priors over the degree of misspecification rather than support restrictions and may also wish to impose proper priors over structural parameters.

The remainder of the paper is organized as follows. In Section 2, we formally introduce our setting and develop the main intuition for the formal results. We then provide specific motivating examples and outline algorithms for sampling from the quasi-posterior or simulating from the asymptotic mixture approximation in Section 3. Section 4 presents the main theoretical results. Finally, we illustrate our approach with simulations and empirical applications in Sections 5 and **??**.

Notation. For a vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ and $q > 0$, we denote $|v|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$,$|v|_\infty = \max_{1\le i\le d} |v_i|$, and $\|v\| = |v|_2$. For a matrix $A = (a_{i,j})_{1\le i\le m, 1\le j\le n}$, we denote the minimum and the maximum singular value of A by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively, the max norm by $|A|_{\max} = \max_{i,j} |a_{i,j}|$, the spectral norm by $\|A\| = \sqrt{\lambda_{\max}(A^T A)}$, and the determinant of a square matrix A when $m = n$ by $\det(A)$. For a vector μ and non-negative definite matrix A, define $\|\mu\|_A := \sqrt{\mu^\top A\mu} \ge 0$. For positive semi-definite matrices A, B, we write $A \ge B$ if $A - B$ is positive semi-definite. For two positive number sequences (a_n) and (b_n), we say $a_n = O(b_n)$ or $a_n \lesssim b_n$ (resp. $a_n \asymp b_n$) if there exists $C > 0$ such that $a_n/b_n \le C$ (resp. $1/C \le a_n/b_n \le C$) for all large n, say $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$, and write $a_n \gg b_n$ if $a_n/b_n \to \infty$ as n diverges. Denote the total variation of moments norm of α for a real-valued measurable function g on Θ by $\|g\|_{TVM(\alpha)} = \int_{h\in\Theta}(1 + \|h\|^\alpha)|g(h)|dh$. We use $\propto$ to denote "proportional to".

For $s > 0$ and a random vector X, we say $X \in \mathscr{L}^s$ if $\|X\|_s = [\mathbb{E}(|X|^s)]^{1/s} < \infty$. We set (X_n) and (Y_n) as two sequences of random variables. Write $X_n = O_p(Y_n)$ if for $\forall \epsilon > 0$, there exists $C > 0$ such that $P(|X_n/Y_n| \le C) > 1 - \epsilon$ for all large n, and say $X_n = o_p(Y_n)$ if $X_n/Y_n \to 0$ in probability as $n \to \infty$. We use the subscript p to denote statements with respect to the outer probability $\mathbb{P}^*$ of a given probability $\mathbb{P}$. We use $\to_d$ to denote convergence in distribution.

2. The Approach: Main Ideas

Suppose that we have data stream $\{Z_t\}$ indexed by $t \subseteq 1, 2, \cdots, T$, which can be time, person, or other unit index. The distribution of each datum Z_t can change with t. We also have an economic model indexed by the parameter $\theta \in \Theta$. We judge the plausibility of this model by taking the unit average over expected scores $g(Z_t, \theta)$:

$$g(\theta) = T^{-1} \sum_t \mathrm{E}[g(Z_t, \theta)].$$

We are interested in pairs $\{\theta, \mu\}$ such that,

$$g(\theta) = \mu,$$

where parameter μ is drawn from the distribution F_μ. We call μ the plausibility characteristic. E.g., In various contexts, μ represents the degree of violation of exogeneity or other exclusion restrictions as well as errors of structural models in explaining the moments of the data encoded by g.

The case $\mu = 0$ with F_μ probability 1 corresponds to the standard GMM case, and θ is plausible if and only if $g(\theta) = 0$. The standard GMM assumes that there is a true parameter value θ_0 such that $g(\theta_0) = 0$. So, the true parameter value is trivially plausible in this context. Under point identification, it is the only plausible value. We depart from this story and allow μ to have a more general distribution. We will denote by θ_μ any root of the equation:

$$g(\theta_\mu) = \mu.$$

The case of $\mu = 0$ is not special anymore. In fact, if F_μ is absolutely continuous, then F_μ assigns zero mass to $\mu = 0$.

We can be equally interested in "most" of plausible θ's that correspond to "most" typical values of μ according to F_μ. For example, we can be interested in the set:

$$PS = \{\theta_\mu : \mu \in \Gamma\},$$

for Γ containing a big, say $1 - a$, fraction of plausibility characteristics μ, under F_μ.

We pursue an (approximate) Bayesian approach to inferring these values using the empirical analogs of the scores as data. We also examine the problem from the angle of making decisions, where we want to make good, optimal (in a certain sense) decisions.

To take the previous reasoning to data, we consider the empirical analog of the average score as:

$$\hat{g}(\theta) := \frac{1}{T}\sum_t g(Z_t, \theta).$$

In some problems, the empirical analogs $\hat{g}(\theta)$ may involve preliminary estimation of nuisance parameters, which we explicitly allow for, as long as the normal key approximation (1) applies to such version. Thus, we use empirical moments as data to input into a Bayesian procedure.

We assume that for any matching pair (θ, μ) such that $g(\theta) = \mu$, we have that the empirical deviation $\hat{g}(\theta) - \mu$ is approximately normal:

$$\sqrt{T}(\hat{g}(\theta) - \mu) \approx_d N(0, \Omega_\theta). \tag{1}$$

This is a mild condition, and a wide variety of central limit theorems provide sufficient conditions.

We can, therefore, use this result to judge the plausibility of pairs (θ, μ). Then

$$c_\theta \exp\left(-T\|\hat{g}(\theta) - \mu\|^2_{\Omega_\theta^{-1}}/2\right)$$

is the (approximate) likelihood for the data $\hat{g}(\theta)$ under the hypothesis that $g(\theta) = \mu$. Given this likelihood the Bayesian posterior on pair $(g(\theta), \mu)$ is:

$$p_T(\theta, \mu) \propto c_\theta \exp\left(-T\|\hat{g}(\theta) - \mu\|^2_{\Omega_\theta^{-1}}/2\right) f_T(\mu) p_0(\theta), \tag{2}$$

where $f_T(\mu)$ is the prior density function for plausibility characteristics μ induced by F_μ, and $p_0(\theta)$ is the prior density over θ, with "flat" (constant) prior being an example. In this the joint prior $\pi(\theta, \mu) = f_T(\mu) p_0(\theta)$. Integrating out μ's then gives the posterior distribution for θ's:

$$p_T(\theta) = \int_\Gamma p_T(\theta, \mu) d\mu. \tag{3}$$

Integrating out θ's gives the posterior distribution

$$p_T(\mu) = \int_\Theta p_T(\theta, \mu) d\theta. \tag{4}$$

of plausibility characteristics μ. This posterior is useful to judge the degree of misspecification of the model.

The highest posterior density region PR_T for θ's containing high mass, say $1-\alpha$, then is a credible region that contains plausible values of θ given the data. That is,

$$PR_T = \{\theta \in \Theta : p_T(\theta) > k\}, \text{ such that } \int_C p_T(\theta) d\theta = 1 - \alpha.$$

The limit case PR_∞, occurring when $V_\theta / T \to 0$ due to $T \to \infty$, is the set of plausible parameter values, reflecting both the plausibility model F_μ and the prior information p_0.

For the case of the flat prior, the PR_∞ region coincides with the theoretical plausibility set PS for $\alpha = a$. Otherwise, it can differ in general due to prior giving more or less weight to certain parameter values. For this reason, PR_∞ is our principal target and not the set PS. Of course, under the strong identification scenario discussed below, the prior would play no role when $T = \infty$ due to the localization of inference at a particular point.

The optimal decision in this framework is obtained by minimizing the posterior expected risk, with expectation taken over the parameter values: Given a loss function $\ell(\theta, d)$ that depends on the parameter θ and a decision $d \in \mathscr{D}$ (regarding the parameter value or some derived quantity such as sufficient welfare statistics), the optimal decision then takes the form:

$$\min_{d \in \mathscr{D}} \int \ell(\theta, d) p_T(\theta) d\theta. \tag{5}$$

This is the general framework. In what follows below, we analyze leading cases that admit some further analytical simplification.

Analytical Case. We elaborate on a lead setting, which admits closed-form or near-closed-form solutions. This, in turn, allows us to perform Plausible GMM inference by a simple adaptation of algorithms for standard GMM inference.

Suppose that plausibility characteristic $\mu \in \mathbb{R}^d$ follows a standard normal distribution,

$$F_\mu = N(0, \Lambda / T),$$

where variance Λ / T scales with T. This captures the idea the variance of the plausibility characteristics is comparable in size to the variance of empirical moments. We can refer to this setting of implausibility being local.

A simple form of Λ is a diagonal matrix with λ_k s on the diagonal, where small values of λ_k indicate that there is little uncertainty about the plausibility of the k-th moment, and the larger value indicates high uncertainty.

We show that this implausibility model results in the following form of the posterior,

$$p_T(\theta) \propto c_\theta \exp(-T\|\hat{g}(\theta)\|^2_{A_\theta}/2)\, p_0(\theta),$$

where

$$A_\theta = \Omega_\theta^{-1} - \Omega_\theta^{-1}[\Lambda^{-1} + \Omega_\theta^{-1}]^{-1}\Omega_\theta^{-1}.$$

We note that the weighting matrix A_θ is different from the weighting matrix Ω_θ^{-1} we have in the standard GMM: The former reflects additional uncertainty brought by variation of μ, that is, plausibility uncertainty. Of course, the extreme case of the standard GMM without plausibility uncertainty is recovered by letting $\Lambda \to 0$ in the formula above.

We next consider the strongly identified case maintaining the local plausibility condition above. The key regularity condition, is that $g(\theta) = \mu$ has the unique solution θ_μ for each μ, and that the following linearization around the pseudo-true value θ_0 holds:

$$g(\theta) = G(\theta - \theta_0) + o(\|\theta - \theta_0)\|),$$

with $G^\top G$ having minimal eigenvalue bounded away from zero. Another condition is the smoothness of the prior around the true value.

Under these conditions, we show that

$$p_T(\theta) \approx \bar{p}_T(\theta) \propto \exp(-T\|\hat{g}(\hat{\theta}) + G(\theta - \hat{\theta})\|^2_{A_\theta}/2),$$

where $\hat{\theta}$ is the posterior mode, which is the GMM estimator with the weighting matrix A_θ. Thus, the approximating posterior has the representation:

$$N(\hat{\theta}, V/T), \quad V = (GA_{\theta_0}G^\top)^{-1}.$$

The variance matrix $V = (GA_{\theta_0}G^\top)^{-1}$ here is different than the GMM variance matrix $(G\Omega_{\theta_0}^{-1}G^\top)^{-1}$, and since $V \geq (G\Omega_{\theta_0}^{-1}G^\top)^{-1}$, the variance matrix V characterizes the additional uncertainty brought by the plausibility considerations. Moreover, the posterior is not centered around the conventional GMM estimator; it is centered around the GMM estimator that uses the weighting matrix A_θ instead of Ω_θ. The latter estimator obeys

$$\hat{\theta} \approx_d N(\theta_0, \bar{V}/T), \quad \bar{V} = (GA_{\theta_0}G^\top)^{-1}GA_{\theta_0}\Omega_{\theta_0}A_{\theta_0}G^\top(GA_{\theta_0}G^\top)^{-1}.$$

It is very interesting to note $\bar{V} \leq V$ because $A_{\theta_0}\Omega_{\theta_0}A_{\theta_0} \leq \Omega_{\theta_0}^{-1}$. That is, variance of the posterior is greater than the variance of the A_{θ_0}-weighted GMM estimator. This means that even though the posterior is centered at the A_{θ_0} weighted GMM, the posterior assigns probability distribution over plausible values of θ, with posterior variance reflecting the variance of plausible values θ that are compatible with data. While these differences may appear to be surprising, this makes sense because the variance of A_{θ_0}- weighted GMM quantifies the uncertainty about θ_0, which is not what the Bayesian posterior does.

For the assessment of the quasi-posterior in the absence of an explicit closed-form solution, we incorporate the Metropolis-Hastings algorithm. This enables the generation of randomized instances of pairs $(\theta^{(i)}, \mu^{(i)})$ drawn from the quasi-posterior, with consideration of potential model inaccuracies. Algorithm 1 delineates the methodology in pseudocode form. This algorithm bears resemblance to the Markov Chain Monte Carlo (MCMC) algorithm investigated by Chernozhukov and Hong (2003). The ease of implementation is complemented by its capacity to accommodate the misspecification term. To minimize the impact of initial value selections, we consider burning periods $n' < n$, and subsequently retain a sequence of generated draws $(\theta^{(n')}, \mu^{(n')}), \cdots, (\theta^{(n)}, \mu^{(n)})$.

3. Examples

In this section, we visit two motivating examples. The first example involves linear moments, while the second example deals with a non-linear and non-smooth quantile moment function.

Example I (Linear IV model). Conley et al. (2012) and Armstrong and Kolesár (2021) discuss the linear IV model with potential model misspecifications.

Conley et al. (2012) alleviate the exclusion restriction and contemplate plausible exogenous instrumental variables. To this end, they present a parameter γ within the first stage regression as a quantifiable gauge of the validity of the exclusion restriction, as illustrated in the following set of equations:

$$Y = X\theta + Z\gamma + \varepsilon; \quad X = Z\Pi + V; \tag{6}$$

where Y represents an $N \times 1$ vector of outcomes; X refers to an $N \times s$ matrix of endogenous variables, with $E[X\varepsilon] \neq 0$, and treatment parameter of interest θ; Z corresponds to an $N \times r$ matrix of instruments where $r \geq s$ with $E\left[Z'\varepsilon\right] = 0$; Π symbolizes a matrix of first-stage coefficients; and γ is the parameter that measures the plausibility of the exclusion restriction.

The sample moments tied to the Ordinary Least Squares (OLS) and Two-Stage Least Squares (2SLS) are expressed as $\hat{g}(\theta) = \frac{1}{T}\sum_{i=1}^{T} X_i(y_i - X_i^\top\theta)$ and $\hat{g}(\theta) = \frac{1}{T}\sum_{i=1}^{T} Z_i(y_i - X_i^\top\theta)$ respectively. Therefore, the conventional OLS/IV-2SLS estimator maximizes the following objective function for a given misspecification term $\mu = 0$:

$$Q(\theta, \mu) = -T\left(\hat{g}(\theta) - \mu\right)^\top W_T(\theta(\mu))\left(\hat{g}(\theta) - \mu\right), \tag{7}$$

Input:

- $p_T(\theta,\mu)$: target distribution that is proportional to $\pi(\theta,\mu)\exp\left\{-\frac{1}{2}Q_T(\theta,\mu)\right\}$;
- $q(\theta',\mu'|\theta,\mu)$: proposal distribution which is a prescribed conditional density;
- (θ^0,μ^0): initial values;
- n: number of iterations.

E.g., our simulations use the estimated pseudo-true value for $\theta^{(0)}$, the prior mode for $\mu^{(0)}$, and $q(\theta',\mu'|\theta,\mu) \propto q_\theta(\theta'|\theta)c(\mu)$ with $q_\theta(\theta'|\theta)$ being the distribution.

Output:

- Samples $\{(\theta^{(i)},\mu^{(i)})\}$ drawn from the target distribution $p_T(\theta,\mu)$.

for $i \leftarrow 1$ **to** n **do**

i. Sample (θ',μ') from $q(\theta',\mu'|\theta^{(i-1)},\mu^{(i-1)})$;

ii. Calculate acceptance ratio

$$\alpha^{(i)} = \min\left(1, \frac{p(\theta',\mu')q(\theta^{(i-1)},\mu^{(i-1)}|\theta',\mu')}{p(\theta^{(i-1)},\mu^{(i-1)})q(\theta',\mu'|\theta^{(i-1)},\mu^{(i-1)})}\right);$$

iii. Sample $u^{(i)}$ from Uniform(0, 1);

if $u^{(i)} \le \alpha^{(i)}$ **then**

Set $(\theta^{(i)},\mu^{(i)}) \leftarrow (\theta',\mu')$;

end

else if $u^{(i)} > \alpha^{(i)}$ **then**

Set $(\theta^{(i)},\mu^{(i)}) \leftarrow (\theta^{(i-1)},\mu^{(i-1)})$;

end

end

Algorithm 1: Metropolis-Hastings Algorithm (MCMC)

where the weighting matrix $W_T(\theta(\mu))$ is given by the identity matrix. Additional choices of the weighting matrix could potentially lead to other estimators, e.g., the Instrumental Variable Generalized Method of Moments (IV-GMM) or the Continuously Updated Estimator (CUE).

When we consider a local Gaussian prior for the misspecification term μ, i.e., $f_T(\mu) \propto \exp\left(-\mu^\top(\Lambda/T)^{-1}\mu/2\right)$, and flat priors for θ, then the quasi-posterior for the θ is easy to compute as $p_T(\theta,\mu) \propto \exp\left(Q(\theta,\mu)/2\right) f_T(\mu)$. Thus, if we assume $W(\theta(\mu)) = W$integrate out μ, we would have

$$p_T(\theta) \propto \exp(-T\|\hat{g}(\theta)\|^2_{W-W[\Lambda^{-1}+W]^{-1}W}/2)$$

Example II (Nonlinear instrumental variable quantile regression (IVQR)). Chernozhukov and Hansen (2004, 2005, 2006) propose the IVQR, which effectively estimates the treatment effects at

various quantiles via instrumental variable regressions. This approach is empirically appealing and has been used in many recent studies, e.g., Glaeser et al. (2015).

Chernozhukov and Hong (2003) discuss one IVQR example for the τth quantile, where they maximize a standard nonlinear GMM objective function (with $\mu = 0$):

$$Q(\theta, \mu) = -T\left(\hat{g}(\theta) - \mu\right)^{\top} W_T(\theta(\mu))\left(\hat{g}(\theta) - \mu\right). \tag{8}$$

In this instance,

$$\hat{g}(\theta) = \frac{1}{T}\sum_{i=1}^{T} m_i(\theta), \quad m_i(\theta) = \left(\tau - 1\left(Y_i \leqslant q\left(D_i, X_i, \theta\right)\right)\right) Z_i, \tag{9}$$

where Y_i represents the scalar dependent variable, D_i is a $d \times 1$ vector of potential endogenous variables, X_i is a $k \times 1$ vector of regressors, Z_i is a $r \times 1$ vector of instruments, and $W_n(\theta)$ is a positive definite weighting matrix, for instance, one specified in Chernozhukov and Hong (2003): $W_n(\theta) = \frac{1}{\tau(1-\tau)}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right]^{-1}$.

The IVQR methodology (see, e.g., Chernozhukov and Hansen (2004)) requests the validity of the IVs but also has an assumption of rank invariance (or rank similarity); these conditions are sensitive to the correct specification of the model. The allowance of model misspecification would robustify the procedure, and the quasi-posterior is easy to implement via the algorithm above.

Remark: The examples mentioned above resort to the conventional models when the misspecification term is held constant at zero. Our interest, however, lies in the outcomes achieved when the values of μ are adjusted to permit a degree of model misspecification.

4. Theoretical results

This section provides theoretical results related to the procedure. We present in Section 4.1 a simple version of the posterior distribution which integrates out μ. Following that, in Section 4.2, we show the main theorem.

4.1. **Simple Proof.** Let us start with a simple proof following Chernozhukov and Hong (2003).

Since we assume a Gaussian prior on μ in this section, we integrate out μ, and then obtain a posterior density $p_T(\theta)$ which will be solely a density function of θ. It shall be noted that Λ plays a vital role in the density of $p_T(\theta)$. We shall regard $W(\theta)$ as a fixed weight matrix. Recall that μ takes value on a compact set Γ. The method of moment estimators involves maximizing

an objective function like the following $Q_T(\theta,\mu)$,

$$\hat{g}(.) = \frac{1}{T}\sum_{t=1}^{T} g(Z_t,\cdot), \tag{10}$$

$$Q_T(\theta,\mu) = -T\left(\hat{g}(\theta)-\mu\right)^{\top} W_T(\theta(\mu))\left(\hat{g}(\theta)-\mu\right), \tag{11}$$

$$W_T(\theta) = W(\theta) + o_p(1) \text{ uniformly in } \theta \in \Theta, \tag{12}$$

$$W(\theta) > 0 \text{ and continuous uniformly in } \theta \in \Theta, \mu \in \Gamma. \tag{13}$$

We denote $p_T(\theta,\mu) = \frac{\pi(\theta,\mu)\exp\{\frac{1}{2}Q_T(\theta,\mu)\}}{\int_\Gamma\int_\Theta \pi(\theta,\mu)\exp\{\frac{1}{2}Q_T(\theta,\mu)\}d\theta d\mu}$, and let $p_T(\theta) = \frac{\int_\Gamma \pi(\theta,\mu)\exp\{\frac{1}{2}Q_T(\theta,\mu)\}d\mu}{\int_\Gamma\int_\Theta \pi(\theta,\mu)\exp\{\frac{1}{2}Q_T(\theta,\mu)\}d\theta d\mu}$.

The case where $\Lambda^{-1}\to 0$ as $T\to\infty$ implies that the prior has little information, while the case where $\Lambda^{-1}\to\infty$ suggest that μ should be fixed at μ_0. The set containing $(\mu,\theta(\mu))$ boils down to a singleton.

Assumption 1. *The pseudo true parameter $\theta(\mu)$ belongs to the interior of a compact convex subset Θ of the Euclidean space R^d. The misspecification parameter $\mu\in R^q$ belongs to the interior of a compact convex subset of Γ. For each μ, there exist a pseudo true parameter $\theta(\mu)$ such that $g(\theta(\mu))=\mu$.*

Let $\theta(\mu_0)=\theta_0$.

Assumption 2. *(Penalty function). $\hat{\theta}$ is the GMM estimator using the weight $W(\theta(\mu))$ which has the expansion $\hat{\theta} = \theta(\mu_0) + J_T\left(\theta(\mu_0)\right)^{-1}\Delta_T\left(\theta(\mu_0)\right) + o_p(1/\sqrt{T})$. (iv) the prior function $\pi:\Theta,\Gamma\to\mathbb{R}_+$ is a continuous, uniformly positive density function. $\pi(\theta,\mu)$ is continuous Θ,Γ, and $\pi(\theta,\mu)$ has positive mass on $\mu\in\Gamma$ for any $\theta\in\Theta$. Let $\pi(\mu,\theta)=\pi(\mu)\pi(\theta|\mu)$, where $\pi(\mu)$ is a Gaussian prior centered at μ_0. It shall be noted that we can assume that $\pi(\theta|\mu)$ is bounded and continuous differentiable around a compact support of θ_0. So it hold that for any positive constant c_0 we have that $\int_{-c_0}^{c_0}\pi(\theta|\mu)d\theta\approx c_0$. and without loss of generality, for example $\pi(\theta|\mu)$ can be a flat prior.*

Remark 1. We shall note that we can switch $\pi(\theta|\mu)$ to some continuous prior on a compact.

Assumption 3. *i) $J_T\left(\theta(\mu)\right)/T>0$ (uniformly over $\theta\in\Theta$ and >0 means positive definiteness) and is continuous, $G(\theta)=\nabla_\theta \mathbb{E}\, g(Z_t,\theta)|_{\theta=\theta(\mu_0)}$ is continuous and full rank, (ii) $\Delta_T\left(\theta(\mu),\mu\right)/\sqrt{T} = -\sqrt{T}(\hat{g}\left(\theta(\mu)\right)-\mu)W(\theta(\mu))G\left(\theta(\mu)\right) \to_{\mathrm{d}} N\left(0,\Omega\left(\theta(\mu)\right)\right), \Omega\left(\theta(\mu)\right)\equiv G\left(\theta(\mu)\right)^{\top}W(\theta(\mu))G\left(\theta(\mu)\right)$.*

Assumption 4. *$r_T(g,\theta) = \sqrt{T}\left|\left(\hat{g}(\theta)-\hat{g}(\theta_0)\right)-\left(\mathbb{E}\hat{g}(\theta)-\mathbb{E}\hat{g}(\theta_0)\right)\right|$. We have that*

$$sup_{\theta:\|\theta-\theta_0\|\le\delta} r_T(g,\theta)/\left([1\vee\sqrt{T}\|\theta-\theta_0\|]\right) = r(\delta),$$

and $r(\delta)\to_p 0$ if $\delta\to 0$. $M_T/\log T\to 0$ for a slow varying constant M_T.

Remark 2. The above Assumption is a type of modulus of continuity Assumption.

Denote

$$N_T(\theta,\mu_0) = \exp\left\{-\frac{1}{2}[V_T(\theta,\mu_0)]\right\} / \int_\Theta \exp\left\{-\frac{1}{2}[V_T(\theta,\mu_0)]\right\}d\theta,$$

where

$$V_T(\theta,\mu) = -2Th_{\theta(\mu)}^{\top} G(\theta(\mu))^{\top} W(\theta(\mu))(\hat{g}(\theta(\mu)) - \mu) + Th_{\theta(\mu)}^{\top} G(\theta(\mu))^{\top} (W(\theta(\mu)) + \Lambda^{-1}) G(\theta(\mu)) h_{\theta(\mu)},$$

and $h_{\theta(\mu)} = \theta - \theta(\mu)$.

Then, we have the main theorem, which shows that the density function $p_T(\theta)$ converges in the TVM norm. With slightly abuse of notation, we denote $\tilde{p}_T(\hat{g}(\theta)) = p_T(\theta)$.

Theorem 1. (Convergence in total variation of moments norm). Under Assumptions $1-4$, for any $0 \leqslant \alpha < \infty$,

$$\left\| p_T(\theta) - N_T(\theta,\mu_0) \right\|_{TVM(\alpha)} \equiv \int_{\theta\in\Theta} \left(1 + \|\theta - \theta(\mu_0)\|^{\alpha}\right) \left| p_T(\theta) - N_T(\theta,\mu_0) \right| \mathrm{d}\theta \to_{\mathrm{p}} 0,$$

$$\int_{\theta\in\Theta} \left(1 + \|\theta - \theta(\mu_0)\|^{\alpha}\right) \left| N_T(\theta,\mu_0) - \tilde{p}_T(\hat{g}(\hat{\theta}) - G(\theta)(\hat{\theta} - \theta)) \right| \mathrm{d}\theta \to_{\mathrm{p}} 0.$$

Proof. See Appendix 7.1. □

Remark 3. We shall note that when $\mu_0 = 0$, and $\Lambda \to \infty$, it boils down to the case of Chernozhukov and Hansen (2006). Theorem 1 shows that $p_T(\theta)$ is concentrated at a $1/\sqrt{T}$ neighborhood of θ_0 as measured by the total variation of moments norm. For large T, $p_T(\theta)$ is approximately a random normal density with random mean parameter $\theta_0 + J_T(\theta_0)^{-1} \Delta_T(\theta_0)/T$, and constant variance parameter $J_T(\theta_0)^{-1}/T$.

Theorem 1, in particular, implies the Bernstein-Von Mises theorems by setting $\alpha = 0$ and $\Lambda \to \infty$, which state the convergence of the likelihood posterior to the limit random density in the total variation norm. Different from Chernozhukov and Hansen (2006), using a Gaussian prior, the integrated posterior function $p_T(\theta)$ is affected by the magnitude of Λ and the prior location μ_0. This indicates that the prior information has an influence on the posterior.

4.2. **Proof of Main Theorems.** Let $\theta \in \mathbb{R}^p$ and $g(\theta) \in \mathbb{R}^q$. In this section, we shall allow p and q to grow concerning n. In the previous section, we consider the μ has a Gaussian prior, and therefore we can integrate out μ. In this section, we discuss an extension of the results. The key insight is that, in the limit, we do not have a Gaussian distribution anymore, but instead, we have a Gaussian mixture distribution. It means that conditional on μ, the limit distribution is Gaussian. Also, the prior distribution of μ is allowed to be approximately Gaussian, and it intervenes in the limit distribution. Recall the relevant estimation framework as the method of moment estimators involves maximizing an objective function of the form in (10). For convenience, we assume that there exists a true point μ_0. Let $G(\theta(\mu_0)) = G$ be an $q \times p$ matrix. Let $B_\varepsilon = \{\theta, \mu : \sqrt{T}\|h(\theta,\mu)\| \leq \varepsilon\}$ with $\varepsilon \lesssim \sqrt{p}\log T$. Without loss of generality, $W = V(\theta(\mu_0))^{-1/2}$ and is fixed at μ_0 throughout the proof. Let $\mu \in \Gamma$ and $\theta \in \Theta$ be two compact support. Let $V(h(.),\theta,\mu) = 2Th(\theta,\mu)^{\top} W(\hat{g}(\theta(\mu_0)) - \mu_0) + T[h(\theta,\mu)]^{\top} W[h(\theta,\mu)] = 2h^{\top} A_\nu + h^{\top} B_\nu h$, where we brief $h(\theta,\mu)$ by h and let $A_\nu = TW(\hat{g}(\theta(\mu_0)) - \mu_0)$, and $B_\nu = TW$. Throughout the section, we keep W to be fixed.

Define

$$N_T(\theta,\mu)=\exp\left\{-\frac{1}{2}[V(h(.),\theta,\mu)]\right\}c(\mu)/\int_{\mu\in\Gamma}\int_{\theta\in\Theta}\exp\left\{-\frac{1}{2}[V(h(.),\theta,\mu)]\right\}c(\mu)d\theta\,d\mu,$$

where $c(\mu)=\exp(-2^{-1}(\mu-\mu_0)^\top\Lambda^{-1}T(\mu-\mu_0))$.

Let $h(\theta,\mu)=G(\theta-\theta(\mu_0))-\mu+\mu_0$. Without loss of generality we can let $\mu_0=0$ and $\theta(\mu_0)=\theta_0$. The $\varepsilon>0$ set is defined by

$$B_\varepsilon=\{\theta,\mu:\sqrt{T}\|h(\theta,\mu)\|\le\varepsilon,\|\mu-\mu_0\|_{\Lambda/T}\le\varepsilon,\theta\in\Theta,\mu\in\Gamma\}.$$

$g(\theta(\mu_0))-\mu_0-g(\theta(\mu))+\mu=0$ by definition. We shall assume that, $G(\theta(\mu_0))\theta(\mu_0)-G(\theta(\mu))\theta(\mu)-\mu_0+\mu\approx 0$. Define the ε set expansion as $h(\theta,\mu)=G(\theta-\theta(\mu_0))-\mu+\mu_0=G(\theta-\theta(\mu_0))-\mu+\mu_0=G\theta-G\theta(\mu_0)-\mu+\mu_0\approx G\theta-G(\theta(\mu))\theta(\mu)$, so $\|h(\theta,\mu)\|\le\varepsilon$ means that $\|G\theta-G(\theta(\mu))\theta(\mu)\|\le\varepsilon$ with $\varepsilon>0$. $\varepsilon\approx c\sqrt{p}\log T$ for a positive constant c. It shall be noted that it automatically includes $\theta(\mu)$ and μ if we assume that for each $\mu\in\Gamma$ there exist a $\theta(\mu)\in\Theta$.

Remark 4. Since the misspecification of the moment is translated into a lack of identification due to overdramatized (θ,μ). First of all, if Λ is a constant matrix with both bounded minimum and maximum eigenvalue. The prior of $\pi(\mu)$ plays a role in the posterior distribution. Thus the set B_ε corresponds to a "local" misspecification case. If $\Lambda^{-1}\to 0$, then it corresponds to the case where there is not so much information from the prior of μ. Then B_ε becomes a $\varepsilon-$ set around the $\Gamma\times\Theta$, and the parameter (θ,μ) shall jointly concentrated around the pseudo true value $(\mu,\theta(\mu))$ with $\mu\in\Gamma$. If $\Lambda^{-1}\to\infty$, then it corresponds to the case of strong information on μ with $\mu=\mu_0$. Then we have the usual case as in Chernozhukov and Hong (2003).

Remark 5. In the high dimensional p and q regime, our theorem shall include both of the two cases. In one case corresponding to severe overidentification, we have $q\gg p$. For example, we have p as a fixed value and q to be of a growing dimension, then it make more sense to assume that $\|G\|\lesssim C, C>0$. Moreover, if we have both p and q growing in the same order, say $p\le q$, but $p\approx q$, this will be a different story.

Assumption 5. *Assume that $\pi(\mu)$ corresponding to the prior density is uni-modal and symmetric around μ_0. Assume the expansion holds uniformly over μ, $\sup_{\mu:\|\mu-\mu_0\|_{\Lambda/T}\le\varepsilon}|\log(\pi(\mu))+2^{-1}(\mu-\mu_0)^\top\Lambda^{-1}T(\mu-\mu_0))|/(\|\mu-\mu_0\|^2_{\Lambda/T}\vee p)\lesssim\sqrt{p}(\log T)/\sqrt{T}$. It shall be noted that we can assume that $\pi(\theta|\mu)$ is bounded and continuously differentiable around a compact support of θ_0. So it hold that for any positive constant c_0 we have that $\int_{-c_0}^{c_0}\pi(\theta|\mu)d\theta\approx c_0$ uniformly over $\mu\in\Gamma$. $\sup_{\mu\in\Gamma}|\pi(\theta(\mu)|\mu)-\pi(\theta(\mu)|\mu_0)|\lesssim\varepsilon/\sqrt{T}$. Without loss of generality, for example, $\pi(\theta|\mu)$ can be a flat prior.*

$$\begin{aligned}\log(\pi(\mu))&=\log(\pi(\mu_0))+\nabla_\mu\pi(\mu_0)\pi(\mu_0)^{-1}(\mu-\mu_0)-2^{-1}(\mu-\mu_0)^\top\Lambda^{-1}T(\mu-\mu_0)+o_p(\|\mu-\mu_0\|_{\Lambda/T}\vee p)\\&=-2^{-1}(\mu-\mu_0)^\top\Lambda^{-1}T(\mu-\mu_0)+o_p(\|\mu-\mu_0\|^2_{\Lambda/T}\vee p),\end{aligned}$$

where the linear term $\nabla_\mu\pi(\mu_0)\pi(\mu_0)^{-1}(\mu-\mu_0)\approx 0$.

Assumption 6. *$tr\{TG^\top(\hat{g}(\theta(\mu_0))-\mu_0)^\top W(\hat{g}(\theta(\mu_0))-\mu_0)G\} = O_p(p)$, and $\lambda_{\max}(WG^\top GW) = O(1)$. $tr\{G^\top W^2 G\} = O(p)$. For each θ exist a pseudo true value $\theta(\mu) \in B_\varepsilon$ such that $\|\theta-\theta(\mu)\| \le \varepsilon$ and $\|G(\theta(\mu)-\theta(\mu_0))-\mu+\mu_0\| = o(\varepsilon^2)$.*

This presumption establishes an identification Assumption through the imposition of restrictions on G. Hansen et al. (2010) elaborate on nonlinear instrumental variables estimators (NLIV) and impose a rank condition on G, thereby precluding instances of weak identification.

Define $R_T(\theta,\mu) = -\frac{1}{2}Q_T(\theta,\mu)+\frac{1}{2}Q_T(\theta(\mu_0),\mu_0)+V(h(.),\theta,\mu)/2-\log(c(\mu))$. Assume the following uniform rate regarding $R_T(\theta,\mu)$. There exists a positive constant ε_0, such that,

Assumption 7. *$\sup_{\theta,\mu\in B_\varepsilon} T|R_T(\theta,\mu)|/(\|\sqrt{T}h(\theta,\mu)\|^2+p) \lesssim \sqrt{p}(\log T)^2/\sqrt{T} \to 0$ with probability 1. And $(\log T)((\log T)p)^{\alpha+1}/T \to 0$ for $\alpha \ge 2$. In addition outside the ball B_ε, we have the with probability approaching 1,*

$$\sup_{\theta,\mu\in B_\varepsilon^c} TR_T(\theta,\mu) \le -\varepsilon^3/\sqrt{T}.$$

Remark 6. $\sup_{\theta,\mu\in B_\varepsilon} TR_T(\theta,\mu)/(\|\sqrt{T}h(\theta,\mu)\|^2+p) \to 0$ is due to the modulus of continuity and locally small oscillation behavior of the empirical process. Note that this Assumption is stronger than Chernozhukov and Hong (2003) but can be verified for differentiable moment functions.

$$\sup_{\theta,\mu\in B_\varepsilon^c} TR_T(\theta,\mu) \le -\varepsilon^2$$

is due to the identification of the likelihood function.

From now on, we choose $W(\theta) = W$. Let $p_T(\theta,\mu) = \frac{\pi(\theta,\mu)\exp\{-\frac{1}{2}Q_T(\theta,\mu)\}}{\int_\Gamma\int_\Theta \pi(\theta,\mu)\exp\{-\frac{1}{2}Q_T(\theta,\mu)\}d\theta d\mu}$.

Theorem 2. Under Assumptions 5-7, we have,

$$\left\| p_T(\theta,\mu) - N_T(\theta,\mu)\right\|_{TVM(\alpha)} \equiv \int_{\mu\in\Gamma}\int_{\theta\in\Theta}\left(1+\|\theta-\theta(\mu)\|^\alpha\right)\left|p_T(\theta,\mu)-N_T(\theta,\mu)\right|\mathrm{d}\theta\, d\mu \to_{\mathrm{p}} 0. \quad (14)$$

Proof. See Appendix 7.2 □

Remark 7. In contrast to Theorem 1, we generalize the prior assumption over μ and do not assume that μ follows a Gaussian prior. But we still allow the variance of μ to be small in the sense that it has enough information to influence the posterior. The limit distribution corresponding to $N_T(\theta,\mu)$ is not a Gaussian density. However, conditioning on μ, the density is Gaussian. Hence, the above Theorem has confidence interval interpretations for any given value of μ. In practice, we can also consider a subset of values of μ and take the union of the confidence intervals with respect to different μ. This corresponds to the procedure in Conley et al. (2012). The coverage would always be larger than those with a fixed μ.

5. Simulation exercises and empirical applications

5.1. **Simulation exercises.** This section continues with the examples described in Section 3 to provide simulation results illustrating the performance in linear and nonlinear moment setups.

5.1.1. *Linear cases.*

Example I continued. The confidence interval constructed based on results from Armstrong and Kolesár (2021), denoted as [AK]-CI in later discussions, and the confidence interval from a Quasi-Bayes approach using a simulation exercise. For simplicity, we consider the case with $s = 1$, $r = 1$, and the sample moments for the IV estimator is

$$\hat{g}(\beta) = Z^\top (Y - X\beta)/T, \tag{15}$$

where for a given γ, $g(\beta) = \mathbb{E}_\gamma \hat{g}(\beta) = \gamma \mathbb{E} Z^\top Z/T$. In the simulation exercises, we calibrate to the 401(K) data employed in Conley et al. (2012). We maintain the same outcome of interest, the net financial assets (1991 dollars, denoted by Y). We consider one endogenous variable, an indicator for 401(k) participation (X), and one instrument variable, an indicator for 401(k) plan eligibility (Z), by first demeaning and projecting out all the other exogenous variables. We compare the coverage rate of the confidence intervals resulting from the following two cases via simulation exercises:

(1) (Quasi-)Bayes approach with the local to zero approximation prior:

$$\gamma|\beta \sim N(0, \delta^2),$$
$$\hat{\beta} \sim_a N(\beta, V_{2SLS} + A\delta^2 A),$$

where $\delta = O(1/\sqrt{T})$ should be comparable to the order of V_{2SLS}.

(2) [AK]-CI with the set of the misspecification term to be

$$\mathscr{C} = \{c : \|c\| \le 2\delta \mathbb{E} Z_i^2\},$$

where we choose such a set to be roughly comparable with the local to zero approximation design.

In the simulation exercises, we select a subsample of size N from the data X, Z[1], and keep them fixed for each simulated sample, draw $\tilde{\varepsilon}$ from $N(0, \hat{\sigma}^2 I_N)$, and generate $\tilde{Y}$ with a given value of γ: $\tilde{Y} = X\widehat{\beta}_0 + Z\gamma + \tilde{\varepsilon}$, where $\widehat{\beta}_0$ and $\hat{\sigma}^2$ are calibrated to the 401(K) data by 2SLS regressing Y over X with instrument Z[2]. We consider a sequence of misspecification levels δ, and for each value of δ, we simulate data choosing $\gamma = \delta$ along with other fixed parameters. For a given choice of the misspecification δ (or equivalently in our simulation design, γ), we compute the average rate of the confidence sets not including β_0 resulting from the two procedures above

[1]Figures 1 and 2 use the total sample of X_p, Z_p (the sample variance of X_p, Z_p are 0.1733708 and 0.2007974 respectively, and $\Pi = 0.1399858$).

[2]Throughout the simulation exercises, $\hat{\sigma}$ and $\widehat{\beta}_0$ are fixed at the estimated values 1881.464 and 15009.6612, respectively.

with the confidence level $\alpha = 95\%$. Therefore, the ideal rejection rate should be 5%. All results are calculated over $M = 500$ simulations.

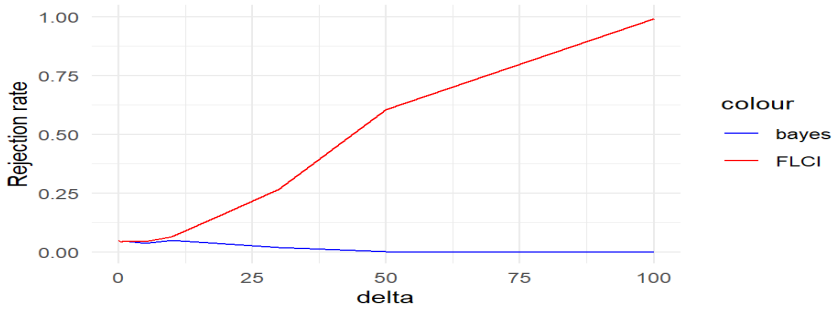

FIGURE 1. Simulated data sample size is 9951 (N=9951, same as the 401(K) sample size used in [CHR]). The red curve is the rejection rate of β_0 using the 95% [AK]-CI, and the blue curve is the rejection rate corresponding to the 95% confidence interval via the Bayes procedure.

Figure 1 presents the rejection rate curve as a function of δ. The range of δ, though large, is comparably reasonable given the value of $\widehat{\sigma}$ or V_{2SLS}. It can be observed that for smaller values of δ, both methods exhibit satisfactory performance. However, as the level of misspecification increases, the [AK]-CI tends to over-reject, while the Bayes procedure maintains the correct coverage, albeit with a tendency to under-reject.

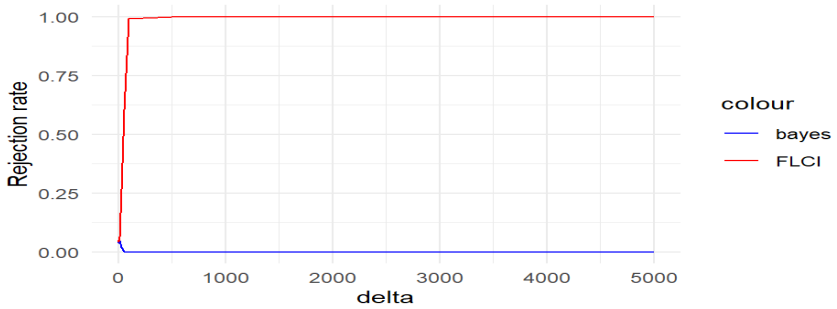

FIGURE 2. Simulated data sample size is 9951 (N=9951, same as the 401(K) sample size used in [CHR]). The red curve is the rejection rate of β_0 using the 95% [AK]-CIm and the blue curve is the rejection rate corresponding to the 95% confidence interval via the Bayes procedure.

Figure 2 plots under the same setting as figure 1 but a more extensive range of δ. The observations above concerning the [AK]-CI, which displays size distortion, may imply that certain asymptotically negligible terms could adversely affect the finite sample performances of the proposed [AK]-method.

5.1.2. *Nonlinear moments.*

Example II continued. We first revisit one Monte Carlo simulation example from Chernozhukov and Hong (2003) with slight modifications to consider our specifications concerning the priors and the potential model misspecification. The Monte Carlo Simulation Example II set up by Chernozhukov and Hong (2003) is:

$$Y_i = \alpha_0 + X_i'\beta_0 + u_i, \quad u_i = \sigma(X_i)\varepsilon_i, \quad \sigma(X_i) = \left(1 + \sum_{j=1}^{3} X_{i,j}\right)/5, \tag{16}$$

where $X_i \sim_{\text{i.i.d}} \exp\mathcal{N}(0, I_3)$ and $\varepsilon_i \sim_{\text{i.i.d}} \mathcal{N}(0,1)$. β's are parameters of interest. They consider the following instrumental moment conditions for the median quantile,

$$g_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2} - 1\left(Y_i \leqslant \alpha + X_i'\beta\right)\right)Z_i, \quad Z_i = (1, X_i^\top)^\top,$$
$$W_n(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2} - 1\left(Y_i \leqslant \alpha + X_i'\beta\right)\right)^2 Z_i Z_i'\right]^{-1}.$$

In the simulation exercise laid out in Table 1, the parameters (α_0, β_0) are equal to the null vector, following the practice in Chernozhukov and Hong (2003). In generating Y_i, u_i is replaced by $\tilde{u}_i = \sigma(D_i)\varepsilon_i + \gamma D_{i,3}^2$ to incorporate some model inaccuracies, with γ assessing the credibility of the instruments Z_i. The mean coverage rate for the true β value within the 2.5% to 97.5% quantiles is displayed in Table 1, as derived from the quasi-Bayesian method without assuming model misspecification (marked as CH, see the quasi-Bayesian approach described in Chernozhukov and Hong (2003)) and the approach we suggest (marked as PGMM). In Table 1, when $\gamma = 0$, both techniques produce comparable outcomes. However, when $\gamma = 1$, CH sometimes results in a strikingly low coverage rate for the actual β_3 value (for example, when $\tau = 0.2$) while incorporating a local implausibility term enhances the coverage rate.

Another critical assumption is the rank invariance (or similarity) utilized in the IVQR with discrete (or bounded continuous) treatment variables, where treatment status should not affect the underlying conditional distribution. We consider the following DGP with potentially missing variables X_i:

$$Y_i = \alpha_0 + D_i^\top\beta_0 + \gamma D_i X_i + \varepsilon_i, \tag{17}$$

where $D_i \sim_{i.i.d.} \text{Bern}(\frac{1}{2})$, $X_i \sim_{\text{i.i.d}} \exp\mathcal{N}(0,0)$ and $(\alpha_0, \beta_0) = (0,1)$. β is a scalar parameter of interest, interpreted as the treatment effect. Suppose the following potentially misspecified instrumental conditions are utilized for estimation of the τ-th quantile,

$$g_n(\theta_\tau) = \frac{1}{n}\sum_{i=1}^{n}\left(\tau - 1\left(Y_i \leqslant \alpha_\tau + D_i'\beta_\tau\right)\right)Z_i, \quad Z_i = (1, D_i^\top)^\top,$$
$$W_n(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2} - 1\left(Y_i \leqslant \alpha + X_i'\beta\right)\right)^2 Z_i Z_i'\right]^{-1}.$$

γ	n	τ	Methods	β_1	β_2	β_3	γ	n	τ	Methods	β_1	β_2	β_3
0	100	0.2	0	0.941	0.947	0.941	0	100	0.2	1	0.995	0.994	0.997
0	100	0.5	0	0.903	0.917	0.901	0	100	0.5	1	0.959	0.969	0.971
0	300	0.2	0	0.917	0.935	0.914	0	300	0.2	1	0.994	0.996	0.993
0	300	0.5	0	0.925	0.919	0.916	0	300	0.5	1	0.965	0.96	0.969
0	700	0.2	0	0.872	0.86	0.873	0	700	0.2	1	0.97	0.975	0.973
0	700	0.5	0	0.922	0.944	0.923	0	700	0.5	1	0.969	0.958	0.959
0	1100	0.2	0	0.875	0.884	0.885	0	1100	0.2	1	0.992	0.981	0.985
0	1100	0.5	0	0.93	0.935	0.933	0	1100	0.5	1	0.962	0.969	0.966
1	100	0.2	0	0.781	0.725	0.273	1	100	0.2	1	0.928	0.929	0.664
1	100	0.5	0	0.554	0.549	0.454	1	100	0.5	1	0.796	0.768	0.74
1	300	0.2	0	0.946	0.941	0.029	1	300	0.2	1	0.985	0.981	0.624
1	300	0.5	0	0.675	0.676	0.34	1	300	0.5	1	0.688	0.67	0.721
1	700	0.2	0	0.974	0.964	0.007	1	700	0.2	1	0.967	0.953	0.628
1	700	0.5	0	0.718	0.723	0.231	1	700	0.5	1	0.65	0.686	0.577
1	1100	0.2	0	0.89	0.907	0.004	1	1100	0.2	1	0.894	0.902	0.617
1	1100	0.5	0	0.773	0.763	0.193	1	1100	0.5	1	0.602	0.604	0.479
10	100	0.2	0	0.518	0.514	0.51	10	100	0.2	1	0.749	0.704	0.75
10	100	0.5	0	0.512	0.498	0.723	10	100	0.5	1	0.774	0.805	0.754
10	300	0.2	0	0.615	0.564	0.385	10	300	0.2	1	0.909	0.879	0.713
10	300	0.5	0	0.45	0.435	0.66	10	300	0.5	1	0.687	0.693	0.713
10	700	0.2	0	0.489	0.498	0.674	10	700	0.2	1	0.671	0.676	0.642
10	700	0.5	0	0.53	0.548	0.678	10	700	0.5	1	0.631	0.666	0.66
10	1100	0.2	0	0.711	0.724	0.321	10	1100	0.2	1	0.883	0.89	0.764
10	1100	0.5	0	0.458	0.413	0.674	10	1100	0.5	1	0.63	0.594	0.45

TABLE 1. The table illustrates the average coverage rate for the true coefficients, symbolized by β, encompassed within the range of the 2.5% and 97.5% quantiles resulted from CH with flat priors over θ's and PGMM with flat priors for θ's and local Gaussian priors $N(0, I)$ for the implausibility term. These mean rates are simulated using the Monte Carlo method, in accordance with the model delineated by (16), involving $\tilde{u}_i$ and a total of 1000 repetitions. The table includes a column marked with n to show the number of simulated samples for each repetition. The character γ symbolizes the extent of misspecification in the model, while the variable τ relates to the specific quantile in the realm of the quantile regressions being examined. The column labeled methods outlines the estimation process, with the value of 0 equating to the CH and the value of 1 signifying the PGMM method.

Then, the missing variable X_i violates the rank invariance (or similarity) as the treatment status after conditioning disturbs the underlying ranks when $\gamma \neq 0$. Table 2 reports similar patterns as observed in Table 1. Additionally, Table 2 shows one column relates to the implausibility level, which shifts away from zero as the misspecification level increases.

γ	n	τ	Methods	β	min c	γ	n	τ	Methods	β	min c
0	100	0.5	-1	0.9666		1	100	0.5	-1	0.0439	
0	100	0.5	0	0.9653		1	100	0.5	0	0.1112	
0	100	0.5	1	0.9807	0.000	1	100	0.5	1	0.4587	0.0000
0	100	0.8	-1	0.9529		1	100	0.8	-1	0.0063	
0	100	0.8	0	0.9455		1	100	0.8	0	0.4694	
0	100	0.8	1	0.9754	0.000	1	100	0.8	1	0.7659	0.0009
0	300	0.5	-1	0.9629		1	300	0.5	-1	0	
0	300	0.5	0	0.9572		1	300	0.5	0	0.0009	
0	300	0.5	1	0.9780	0.000	1	300	0.5	1	0.3555	0.0034
0	300	0.8	-1	0.9573		1	300	0.8	-1	0	
0	300	0.8	0	0.9574		1	300	0.8	0	0.0893	
0	300	0.8	1	0.9796	0.000	1	300	0.8	1	0.7009	0.0003
0	700	0.5	-1	0.9581		1	700	0.5	-1	0	
0	700	0.5	0	0.9505		1	700	0.5	0	0.0002	
0	700	0.5	1	0.9722	0.000	1	700	0.5	1	0.2289	0.18247
0	700	0.8	-1	0.9548		1	700	0.8	-1	0	
0	700	0.8	0	0.9612		1	700	0.8	0	0.0374	
0	700	0.8	1	0.9556	0.0105	1	700	0.8	1	0.7769	0.9083

TABLE 2. The table displays the average coverage rate for the actual coefficients, symbolized by β, within the range of the union of the 2.5% and 97.5% quantiles of the conditional draws given each value of c near their individual quantiles from 2.5% to 97.5% quantiles, resulted from CH with flat priors over θ's and PGMM with flat priors for θ's and local Gaussian priors $N(0, I/\sqrt{n})$ for the implausibility term. This mean rate is based on the Monte Carlo Simulation by the model depicted by (17), conducted over 1000 iterations. In the table, the column marked with n states the magnitude of the sample simulated for each cycle, the symbol γ signifies the extent of mismatch in the model, the parameter τ relates to the intended quantile in the quantile regressions being evaluated, and the column titled methods outlines the method of estimation. The value of -1 equates to the IVQR method, 0 corresponds to the CH, and 1 is linked to the PGMM method. There are two moment equations utilized in this simulation exercise, and the column labeled min c reports the mean value for a specific indicator that is set to 1 if the range between the 2.5% and 97.5% quantiles for the PGMM posteriors for at least one of the two entries of the implausibility term does not include zero.

Figures 3-4 show simulated results from two specific iterations as presented in Table 2. These figures demonstrate similar patterns in the posterior probabilities of the βs for CH and PGMM when $\gamma = 0$, but they exhibit differences when $\gamma = 1$. Furthermore, it is noticeable that when $\gamma = 0$, the posteriors of the implausibility term tend to converge towards zero, which is not the case when $\gamma = 1$. Another intriguing observation can be seen in Figure 4-(a.3)(b.3), where we notice that the posterior masses tend to gather around the true value of β while also exhibiting non-zero values for the implausibility terms.

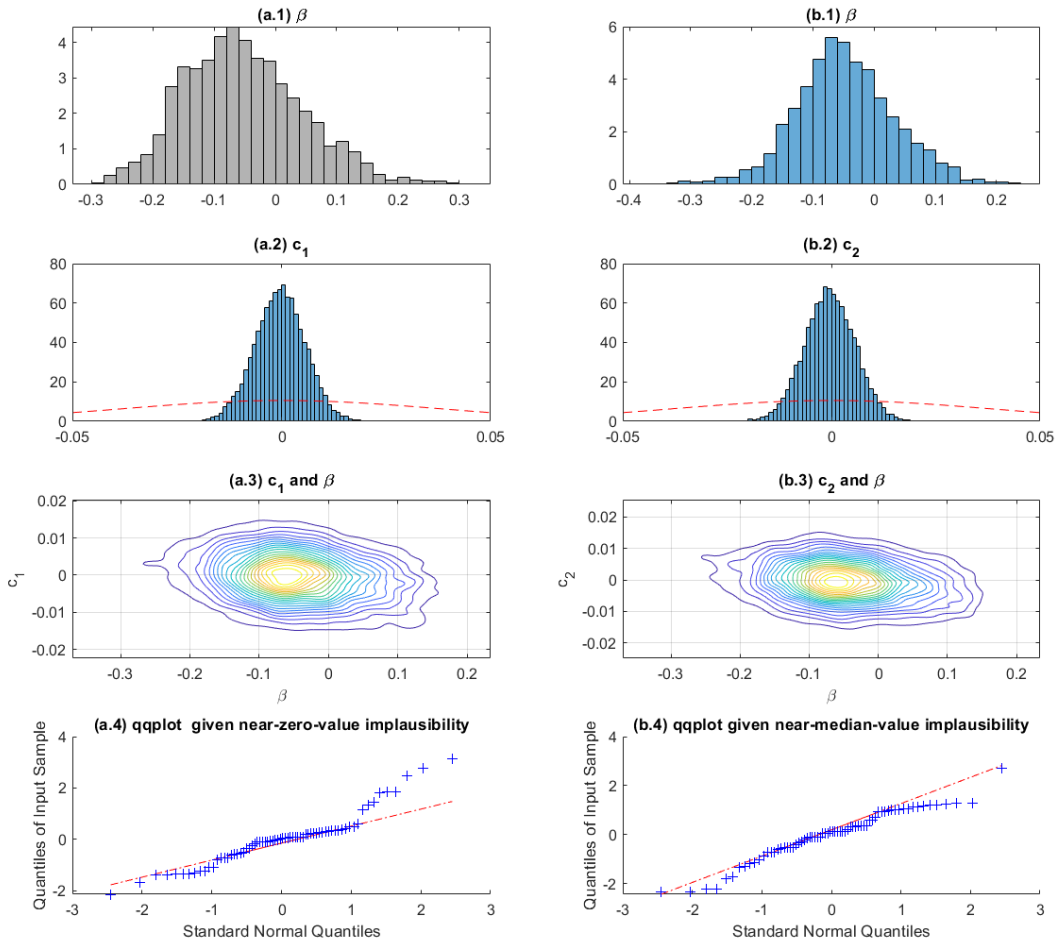

FIGURE 3. (a.1)(a.2) create histograms to display the distribution of simulated values for β from the posterior distributions obtained using CH and PGMM methods, respectively. (a.2)(b.2) construct a histogram to show the distribution of simulated values for the implausibility terms from the PGMM posterior. (a.3)(b.3) generate joint contour plots to illustrate the relationship between the implausibility terms and β values. (a.4)(b.4) plot the QQ-plots of selected β draws corresponding to near-zero/near-median draws of the two misspecification terms. The model used is identical to the one described in Table 2 with $n = 700, \gamma = 0, \tau = 0.5$.

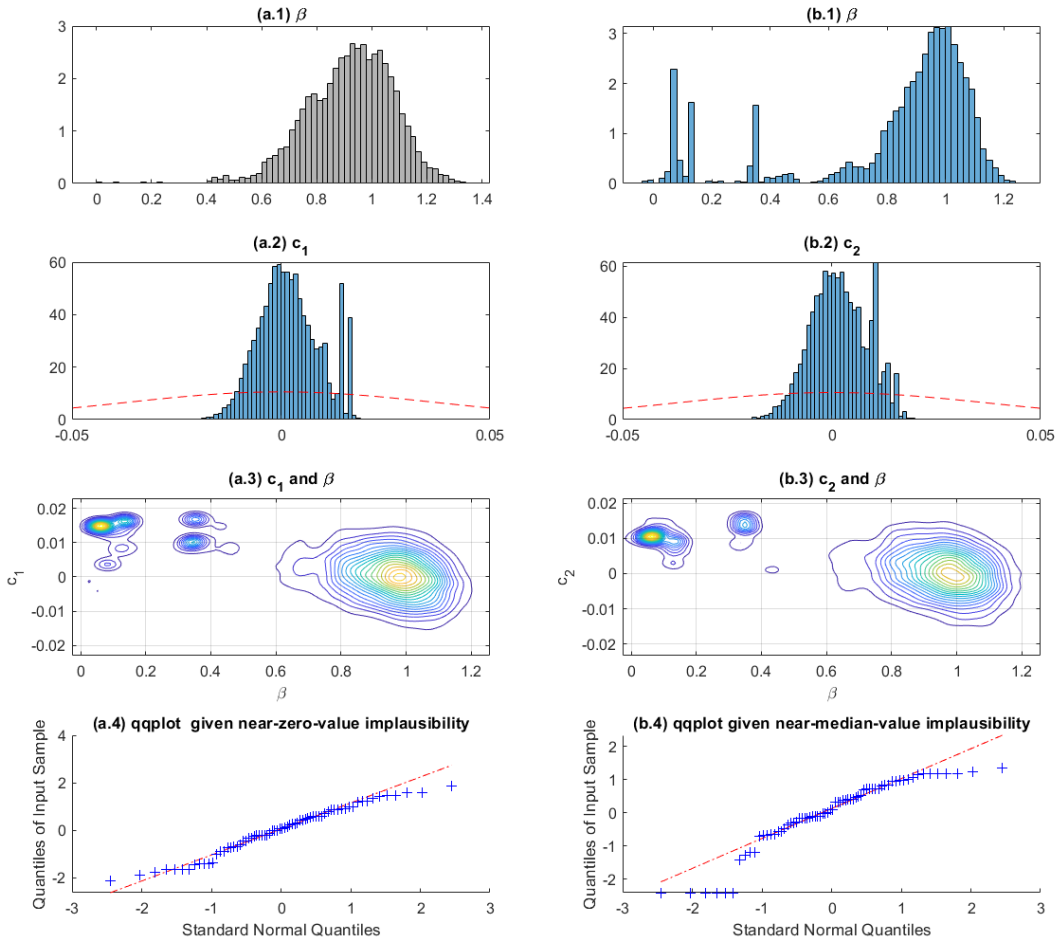

FIGURE 4. Like Figure 3, but with n equal to 700, γ equal to 1, and τ equal to 0.5.

5.2. **Empirical applications.** In this section, we consider similar models as described in Section 5.1 and show that our proposed procedures are of empirical relevance.

5.2.1. *Choice of prior and misspecification level.* In empirical applications, we opt for ad hoc priors for the implausibility term. We employ a local Gaussian prior to simulation exercises. The Gaussian priors are common choices in the literature (e.g., Conley et al. (2012),Spokoiny (2017)).

In the application of the linear IV model, we calculate based on priors with variances proportional to the misspecification level. In the non-linear IVQR setting, we start with the $N(0, I/n)$ prior that does not depend on the moment conditions. Then, we also consider priors that depend on the moment conditions, considering the scale of instrument variables; for example, we consider the following prior to the misspecification level, whose variance would be proportional to the product of the variances of the sample errors at the true value, i.e., $\text{Var}(\hat{g}_i(\theta)) = \text{Var}\left(\tau - 1\left(Y_i \leqslant \alpha_\tau + D_i'\beta_\tau\right)\right) = \varpi - \varpi^2$ with $\varpi = E\left(\tau - 1\left(Y_i \leqslant \alpha_\tau + D_i'\beta_\tau\right)\right)$, pretending as if there was no misspecification, and of the instrument variables employed in the moment conditions, i.e., $\text{Var}(Z_i)$.

5.2.2. *Linear moments application and comparison with [AK].*

Example I continued. As indicated by the simulation exercises, the [AK] method takes into account local violations and may exhibit poorer performance when the prior information suggests significantly larger misspecification levels or in the context of relatively small samples.

Figure 5 displays the confidence intervals generated based on the 401(K) data, along with various misspecification levels (δ's). The length of [AK]-CI exhibits a slower rate of change with increasing δ's.[3] Moreover, these findings may imply that FLCI could over-reject when misspecification levels are relatively larger in finite samples.

5.2.3. *Nonlinear moments application.*

Example II continued. This part contemplates the empirical implementations derived from Autor et al. (2017). This evaluation mirrors Table 4 from Autor et al. (2017), albeit with a simpler version presented in Figure 7-(a.2)(a.3)(b.1) that all the other additional control covariates are excluded during the estimation procedure.

[3]The Bayes procedure can also generate confidence intervals comparable to those of [AK] by carefully choosing data-driven priors. As implied by Figure 5, the associated priors must remain approximately constant as δ increases to obtain similar results using the quasi-Bayes approach.

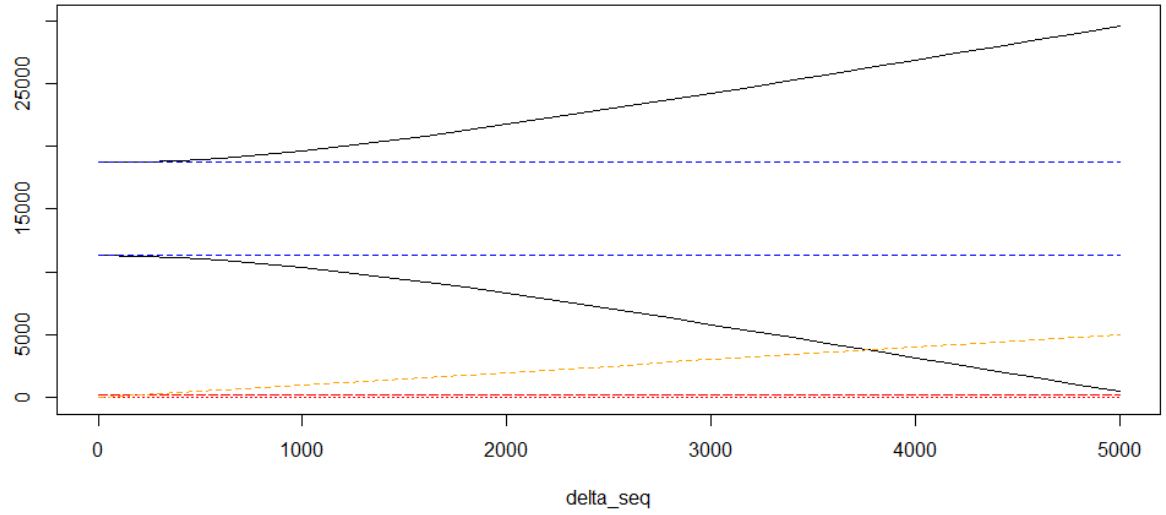

FIGURE 5. Black solid line (Bayes confidence interval), blue dashed line ([AK] confidence interval), red dashed lines (the corresponding $\mu^*(\delta)$ and $\sqrt{W_\mu^*(\delta)}$ such that Bayes procedure with the prior $\gamma|\beta \sim N(\mu^*(\delta), W_\mu^*(\delta))$ would deliver similar confidence intervals as [AK]), orange curve plots the curve $f(\delta) = \delta$.

Figure 7 estimates the relationship across various points between any form of employment (temporary help or direct hire) during the Work First tenure and earnings through the quantile regression, in which participants job placements instrumented by the average excess job placement probabilities of Work First contractors in the year in which the participant entered the Work First program. Comparative analysis is conducted among the estimates derived from the IVQR (see, e.g., Chernozhukov and Hansen (2006), Chernozhukov and Hansen (2005)), CH (incorporating flat priors without assuming model misspecification), and then our proposed method, denoted by PGMM (utilizing flat priors from parameters and a local Gaussian prior for the implausibility term).

By specifications, the models being studied in Figures 6,7-(a.2)-(b.1) may be mischaracterized due to the removal of extra control variables when compared to the model in Figures 6,7-(a.1). It can be observed that the grey sections in (a.2)-(a.3) differ from the pale grey regions, yet when a local implausibility term is applied, the grey regions appear to encompass the pale grey regions, though with a significantly larger bandwidth. The grey sections do not exactly match the pale grey sections in (b.1), particularly around the 50th quantile in Figure 6, and consequently, the associated posterior of the implausibility term c shifts away from the zeros at these specific locations, as depicted, for example, in Figure 6-(b.3).

These observations are consistent with the simulation outcomes mentioned in Section 5.1. Our suggested approach might result in a wider scope if the outcomes are susceptible to minor

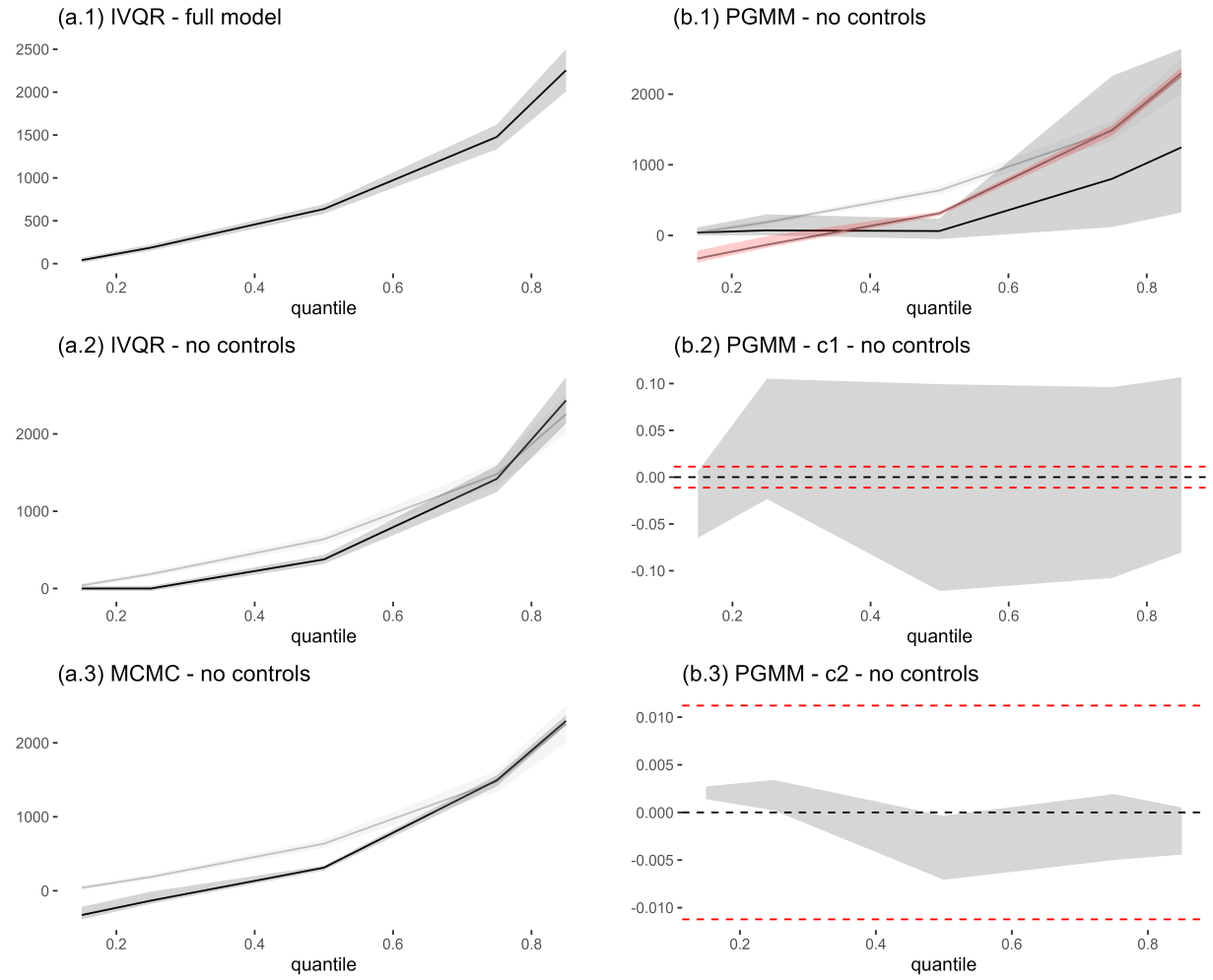

FIGURE 6. Confidence interval/2.5th and 97.5th quantiles concerning the constant parameter along various earnings quantiles. Other explanations pertain to the caption of Figure 7

disturbances, due to which considering the possibility of incorrect model specifications might be a more rational approach, and it can still be informative and insightful.

6. CONCLUSION

We suggest using a quasi-Bayesian method to ease the moment conditions in the GMM framework. This approach is useful when there is a potential model misrepresentation and a violation of null moment conditions. Through simulation exercises and real-world examples, we showcase the characteristics of our proposed techniques and compare them to other established methodologies.

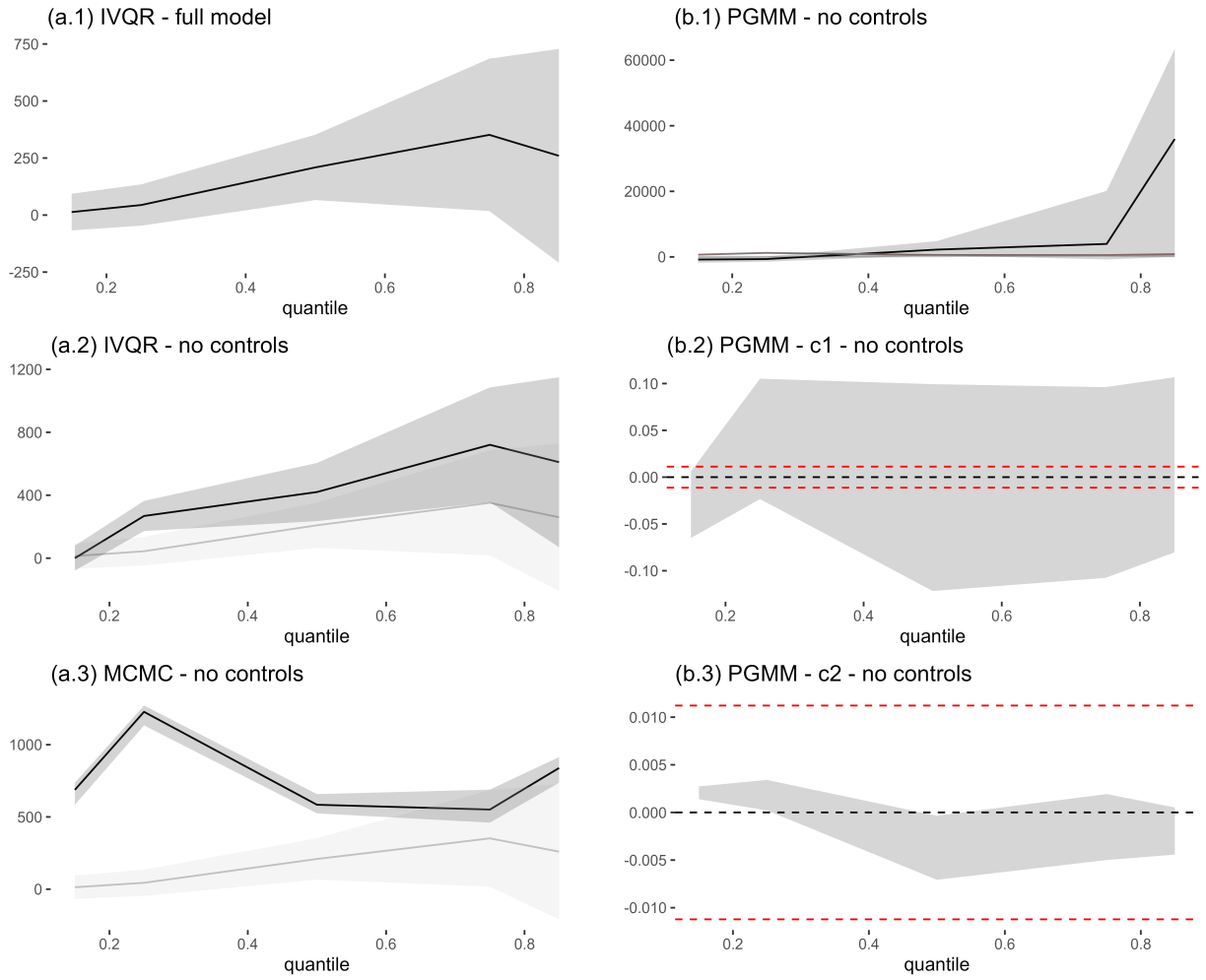

FIGURE 7. Confidence interval/2.5th and 97.5th quantiles of the slope coefficient concerning the variable *any job placement* along various quantiles of earnings. (a.1): A black solid line represents the calculated values from the full model application of the instrumental variable quantile regression method (IVQR), incorporating all control variables as detailed in Table 4 from the cited work by Autor et al. (2017). Surrounding this line, a gray shadowed region highlights the associated 95% confidence intervals, determined through conventional standard errors. (a.2): This representation is similar to (a.1), but in this instance, the IVQR method's full model estimates exclude all control variables other than the endogenous variable *Any job placement*. The gray shadowed region once again denotes the 95% confidence intervals, and the pale grey regions correspond to the shadowed area outlined in (a.1). (a.3) and (b.1): These portions follow the pattern of (a.2), but the gray area here indicates the range between the 2.5th and 97.5th quantiles. These are calculated based on CH with flat priors over the parameters and the proposed PGMM with a prior $N(0, 1/\sqrt{n})$ over the implausibility term c and flat priors for the other parameters. Additionally, the black solid line designates the median that comes from the associated posterior, and the light red shadowed area in (b.1) corresponds to the grey shadowed area in (a.2). Regarding (b.1), two moment condition equations are employed, meaning that the term c includes two entries. In (b.2) and (b.3), the gray shadowed areas pinpoint the corresponding 2.5th and 97.5th quantiles of these two entries from the posterior, and the dashed line designates the median, along with the 2.5th and 97.5th quantiles of their priors.

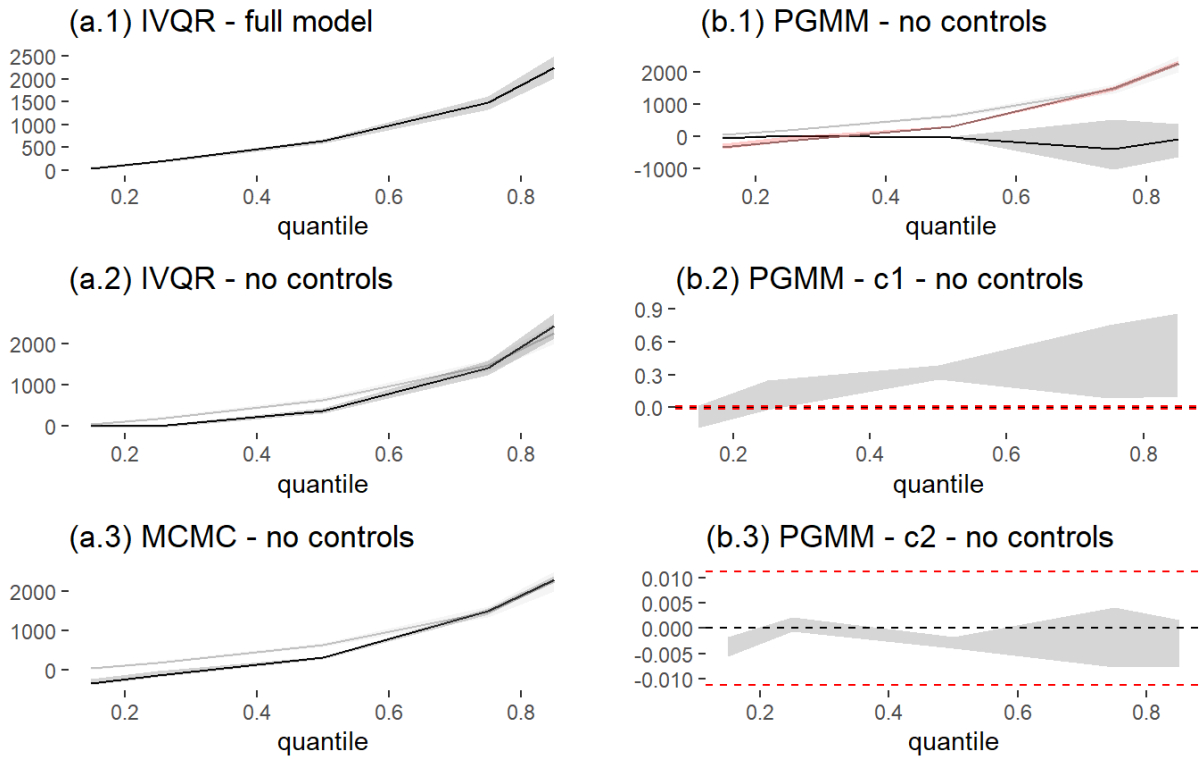

FIGURE 8. Confidence interval/2.5th and 97.5th quantiles concerning the constant parameter along various earnings quantiles. The variance of the prior Other explanations pertain to the caption of Figure 7

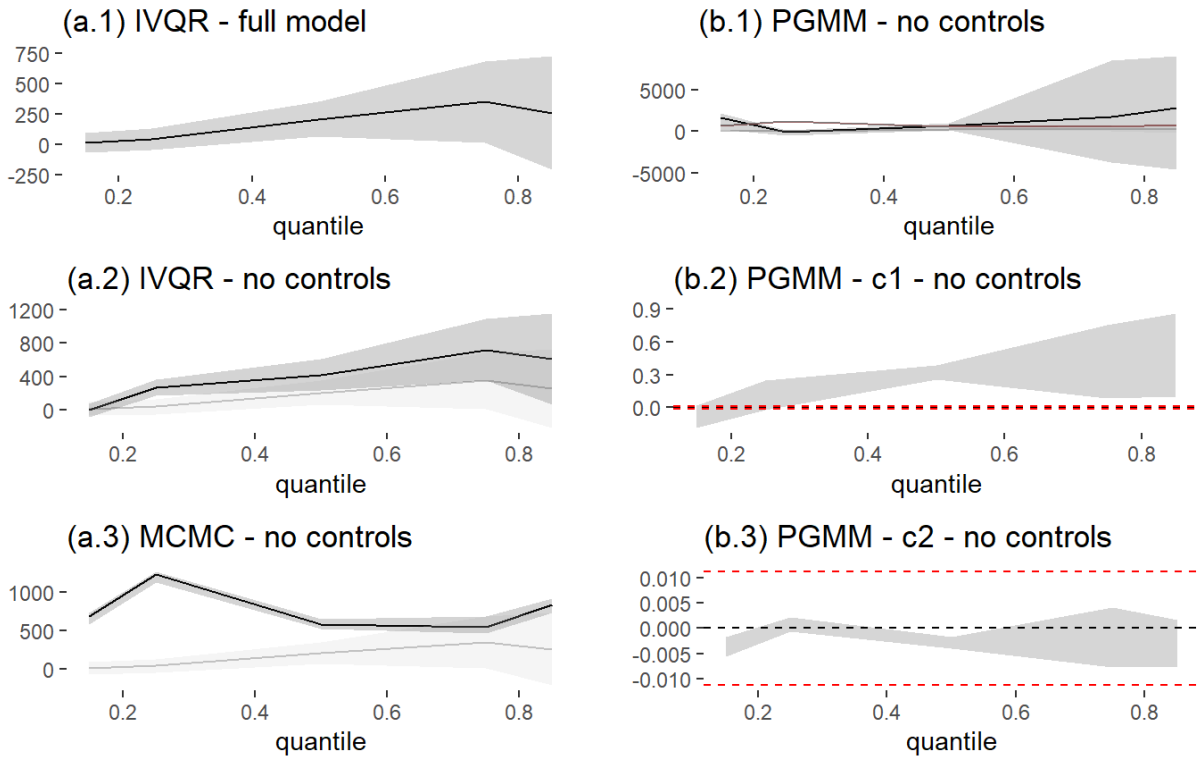

FIGURE 9. Confidence interval/2.5th and 97.5th quantiles concerning the constant parameter along various earnings quantiles. The variance of the prior of the implausibility term is set to be proportional to the product of the variances of the sample errors at the true value, i.e., $\mathrm{Var}(\hat{g}_i(\theta)) = \mathrm{Var}\left(\tau - 1\left(Y_i \leqslant \alpha_\tau + D_i'\beta_\tau\right)\right) = \varpi - \varpi^2$ with $\varpi = E\left(\tau - 1\left(Y_i \leqslant \alpha_\tau + D_i'\beta_\tau\right)\right)$, pretending as if there was no misspecification, and of the instrument variables employed in the moment conditions, i.e., $\mathrm{Var}(Z_i)$. Other explanations pertain to the caption of Figure 7

7. Appendix A

7.1. **Proof of Theorem 1.** The first result is a direct conclusion following Chernozhukov and Hong (2003), Proposition 1, it shall be noted that the $p_T(\theta)$ differs slightly by a different scale $\sqrt{T}$. Now, we show the second statement.

To prove the second result, we define a function $r(\hat{\theta}-\theta)$ such that, $\hat{g}(\hat{\theta})-\hat{g}(\theta)-G(\theta)(\hat{\theta}-\theta) = res(\hat{\theta}-\theta)$. It suffice to prove $res(\hat{\theta}-\theta)$ is small on $\{\theta : \|\theta-\theta_0\| \leq M_T/\sqrt{T}\}$.

So on $\{\theta : \|\theta-\theta_0\| \leq M_T/\sqrt{T}\}$,

$$\hat{g}(\theta) = \hat{g}(\hat{\theta}) - G(\theta)(\hat{\theta}-\theta) - res(\hat{\theta}-\theta).$$

By Assumption 3 i), on $\{\theta : \|\theta-\theta_0\| \leq M_T/\sqrt{T}\}$,

$$\mathbb{E}\hat{g}(\hat{\theta}) - \mathbb{E}\hat{g}(\theta) = G(\theta)(\hat{\theta}-\theta) + o(\|\hat{\theta}-\theta\|).$$

By Assumption 3 i), $\|\hat{\theta}-\theta_0\| \lesssim_p \sqrt{T}^{-1}$. Denote $|.|_a$ as the elementwise absolute value. By Assumption 3 iii),

$$|\hat{g}(\hat{\theta})-\hat{g}(\theta)-\mathbb{E}\hat{g}(\hat{\theta})+\mathbb{E}\hat{g}(\theta)|_a/(1+\sqrt{T}\|\hat{\theta}-\theta\|) \leq \sup_{\{\theta:\|\theta-\theta_0\|\leq M_T/\sqrt{T}\}} |r_T(g,\theta)|_a/[\sqrt{T}(1+\sqrt{T}\|\hat{\theta}-\theta\|)] \to_p 0,$$

since $M_T/\sqrt{T} \to 0$, where M_T is a slow varying term denoted by $\log T$.

So on the ball $\{\theta : \|\theta-\theta_0\| \leq \log T/\sqrt{T}\}$, it holds uniformly over θ, $\|res(\hat{\theta}-\theta)\| \leq \sup_{\{\theta:\|\theta-\theta_0\|\leq M_T/\sqrt{T}\}} \|r_T(g,\theta)\|\sqrt{T}^{-1}(1+\sqrt{T}\|\hat{\theta}-\theta\|)+\|G(\theta)(\hat{\theta}-\theta)\|+\|\mathbb{E}\hat{g}(\hat{\theta})-\mathbb{E}\hat{g}(\theta)\| = O_p(\sqrt{T}^{-1}(1+\sqrt{T}\|\hat{\theta}-\theta\|)) + O_p(\|\hat{\theta}-\theta_0\|) = o_p(1)$. Thus together with the argument in Chernozhukov and Hong (2003), we have

$$\int_{\theta\in\Theta} \left(1+\|\theta-\theta(\mu_0)\|^{\alpha}\right)\left|p_T(\theta) - p_T(\hat{g}(\hat{\theta}) - G(\theta)(\hat{\theta}-\theta))\right| \mathrm{d}\theta \to_{\mathrm{p}} 0.$$

Then, the second conclusion holds.

7.2. **Proof of Theorem 2.** Define $A = G^\top W(\hat{g}(\theta(\mu_0))-\mu_0)$, $B = G^\top WG$ and $C = 2^{-1}(\mu-\mu_0)^\top(\Lambda^{-1})T(\mu-\mu_0)-2\log(\pi(\theta(\mu)|\mu))$.

Then let us analyse $V(h(.),\theta,\mu)$. By Assumption 6, we have, we can replace, $\theta(\mu_0)^\top G^\top - \mu + \mu_0$ with $\theta(\mu)^\top G^\top(\theta(\mu))$. Then we have

$$\begin{aligned} &V(h(.),\theta,\mu) - 2\log\pi(\theta(\mu),\mu) = 2T[(\theta-\theta(\mu))^\top G^\top]W(\hat{g}(\theta(\mu_0)) - \mu_0) \\ &+T[(\theta-\theta(\mu))^\top G^\top]W[G(\theta-\theta(\mu))] - 2\log\pi(\mu)-2\log(\pi(\theta(\mu)|\mu)). \end{aligned} \tag{18}$$

Thus, we reformulate the equation as,

$$
\begin{aligned}
&V(h(.),\theta,\mu) - 2\log\pi(\theta(\mu),\mu) \\
&= 2T[(\theta-\theta(\mu_0))^\top G^\top - \mu^\top + \mu_0^\top]W(\hat{g}(\theta(\mu_0)) - \mu_0) \\
&+ T[(\theta-\theta(\mu_0))^\top G^\top - \mu^\top + \mu_0^\top]W[(\theta-\theta(\mu_0))^\top G^\top - \mu^\top + \mu_0^\top]^\top - 2\log\pi(\mu) - 2\log(\pi(\theta_0|\mu)) \\
&= 2T[(\theta-\theta(\mu))^\top G^\top]W(\hat{g}(\theta(\mu_0)) - \mu_0) \\
&+ T[(\theta-\theta(\mu))^\top G^\top]W[(\theta-\theta(\mu))^\top G^\top]^\top - 2\log\pi(\mu) - 2\log(\pi(\theta(\mu)|\mu)) \\
&= 2T(\theta-\theta(\mu))^\top A + T(\theta-\theta(\mu))^\top B(\theta-\theta(\mu)) + 2C + o_p(\|\mu-\mu_0\|^2_{\Lambda/T} \vee p \vee \varepsilon).
\end{aligned}
$$

We know that this is proportional to the log-likelihood of the density function of $N(B^{-1}A, (TB2)^{-1})$.

By Assumption 5,

$$
\begin{aligned}
tr(AA^\top) &= tr((\hat{g}(\theta(\mu)) - \mu)^\top WGG^\top W(\hat{g}(\theta(\mu)) - \mu)) \\
&= tr(G^\top W\mathbb{E}[(\hat{g}(\theta(\mu)) - \mu)(\hat{g}(\theta(\mu)) - \mu)^\top]WG) \\
&= O_p(T^{-1}p).
\end{aligned}
$$

We shall assume that there exists a constant $C > 0$ such that on the ball Γ, the radius is bounded. By Assumption 6, we shall focus on the ball, $\|\mu-\mu_0\|_{\Lambda/T} \lesssim \lambda_{\max}(\Lambda)(\log T)\sqrt{p}/\sqrt{T} \wedge C$, and $\lambda_{\max}(GG^\top) = \lambda_{\max}(G^\top G) \lesssim_p 1$. One remark is that if $\Lambda^{-1} \to 0$ the density of the prior of μ is flat. Then, we should restrict at compact support of μ to integrate. $\exp(-V(h(.),\theta,\mu)/2)c(\mu)$ for any fixed μ is proportion to multivariate normal with mean $B^{-1}A$ and variance B^{-1}.

To derive the conclusion, we shall divide the proof into the following steps:

$$
\begin{aligned}
&\int_\Gamma \int_\Theta \left(1 + \|\theta-\theta(\mu)\|^\alpha\right) \left| p_T(\theta,\mu) - N_T(\theta,\mu) \right| \mathrm{d}\theta\, d\mu \\
&= \int_{B_\varepsilon} \left(1 + \|\theta-\theta(\mu)\|^\alpha\right) \left| p_T(\theta,\mu) - N_T(\theta,\mu) \right| \mathrm{d}\theta\, d\mu \\
&+ \int_{B_\varepsilon^c} \left(1 + \|\theta-\theta(\mu)\|^\alpha\right) \left| p_T(\theta,\mu) - N_T(\theta,\mu) \right| \mathrm{d}\theta\, d\mu = I_1 + I_2.
\end{aligned}
$$

To look at I_1, we let $\int_{B_\varepsilon} \left(1 + \|\theta-\theta(\mu)\|^\alpha\right) N_T(\theta,\mu) \left| p_T(\theta,\mu)/N_T(\theta,\mu) - 1 \right| \mathrm{d}\theta\, d\mu$. Let the integral ratio be

$$
c(\theta_0,\mu_0) = \frac{\int_{\mu,\theta\in B_\varepsilon} \exp(-V(h(.),\theta,\mu)/2 + \log\pi(\mu,\theta(\mu)))d\theta\, d\mu}{\int_{\mu,\theta\in B_\varepsilon} \exp(-\frac{1}{2}Q_T(\theta,\mu) + \frac{1}{2}Q_T(\theta(\mu_0),\mu_0) + \log\pi(\mu,\theta))d\theta\, d\mu}.
$$

We see that,

$$
\frac{p_T(\theta,\mu)}{N_T(\theta,\mu)} = \frac{\exp\left(-\frac{1}{2}Q_T(\theta,\mu) + \frac{1}{2}Q_T(\theta(\mu_0),\mu_0) + V(h(.),\theta,\mu)/2 - \log(c(\mu))\right)}{c(\theta_0,\mu_0)}.
$$

Let γ be an $p-$ dimensional standard Gaussian distribution. Recall that $\pi(\mu)$ be a function proportional to the density of Gaussian distribution with variance $(\Lambda^{-1}T)^{-1/2}$ and mean μ_0. It is not hard to see that condition on μ, the density function $N_T(\theta|\mu)$ is a density function of a multivariate Gaussian random variable with mean $B^{-1}A$ and variance B.
Following fact about Gaussian Integral, with $\boldsymbol{x} \in \mathbb{R}^n$, we have $\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\boldsymbol{x}^T B\boldsymbol{x} + A^T\boldsymbol{x}\right) dx_1 dx_2 \ldots dx_n = \frac{(2\pi)^{n/2}}{|B|^{1/2}} \exp\left[\frac{1}{2}A^T B^{-1} A\right]$, with $|B|$ denoted as the determinant of a matrix B.

We define $N_T(\theta|\mu) \stackrel{\text{def}}{=} N_T(\theta,\mu)/\pi(\mu)$. And we define $\mathbb{E}_{N_T(\theta|\mu)}(.)$ as taking expectation under $N_T(\theta|\mu)$. Let $a(\theta,\mu) = \exp(R_T(\theta,\mu))/c(\theta_0,\mu_0) - 1$ and $\delta > 0$, then

$$
\begin{aligned}
I_1 &= \int_{B_\varepsilon} \left(1 + \|\theta - \theta(\mu)\|^\alpha\right) N_T(\theta,\mu) \left|\exp(R_T(\theta,\mu))/c(\theta_0,\mu_0) - 1\right| \mathrm{d}\theta\, d\mu \\
&= \int_{B_\varepsilon} \left(1 + \|\theta - \theta(\mu)\|^\alpha\right) N_T(\theta,\mu) \left|a(\theta,\mu)\right| \mathrm{d}\theta\, d\mu \\
&= \int_{B_\varepsilon} \left(1 + \|\theta - \theta(\mu)\|^\alpha\right) N_T(\theta|\mu) \left|a(\theta,\mu)\right| \mathrm{d}\theta \pi(\mu)\, d\mu \\
&\leq \delta (\log T)^3 \sqrt{p^3}/\sqrt{T} \int_{B_\varepsilon} N_T(\theta|\mu) \left|\left(1 + \|\theta - \theta(\mu)\|^\alpha\right)\right| \mathrm{d}\theta \pi(\mu)\, d\mu \\
&\leq \delta (\log T)^3 \sqrt{p^3}/\sqrt{T} \int_{\mu \in \Gamma_\varepsilon} \mathbb{E}_{N_T(\theta|\mu)} (\|\sqrt{T}^{-1} B^{-1/2}\gamma - \sqrt{T} B^{-1/2} A\|^\alpha \\
&\mathbf{1}(\|\sqrt{T}^{-1} B^{-1/2}\gamma - \sqrt{T} B^{-1/2} A)\| \leq \varepsilon)) \pi(\mu)\, d\mu \\
&\lesssim (\log T)^3 \delta (\sqrt{p})^{\alpha+3}/\sqrt{T}.
\end{aligned}
$$

Denote $\mathbb{P}_{N_T(.|.)}$ as the conditional distribution function of $\theta - \theta(\mu)$ conditioning on a fixed value of μ. $\Gamma_\varepsilon = \{\mu : \|\mu - \mu_0\|_{\Lambda/T} \leq \varepsilon\}$.

$$
\begin{aligned}
I_2 &= \int_{B_\varepsilon^c} \left(1 + \|\theta - \theta(\mu)\|^\alpha\right) N_T(\theta,\mu) \left|\exp(R_T(\theta,\mu))/c(\theta_0,\mu_0) - 1\right| \mathrm{d}\theta\, d\mu \\
&\leq c \int_{B_\varepsilon^c} N_T(\theta|\mu) \left|\left(1 + \|\theta - \theta(\mu)\|^\alpha\right)\right| \mathrm{d}\theta \pi(\mu)\, d\mu \\
&\leq \int_{\mu \in \Gamma} \mathbb{P}_{N_T(\theta|\mu)} (\|\theta - \theta(\mu)\|^\alpha > \varepsilon|\mu) \pi(\mu)\, d\mu \\
&\leq \int_{\mu \in \Gamma} \mathbb{P}_\gamma (\|\sqrt{T}^{-1} B^{-1/2}\gamma - \sqrt{T} B^{-1/2} A\|^\alpha > \varepsilon|\mu) \pi(\mu)\, d\mu \\
&\leq \sqrt{p}^\alpha \int_{\Gamma_\varepsilon} \exp(-\varepsilon - \sqrt{p}) \pi(\mu)\, d\mu \to 0,
\end{aligned}
$$

where Γ_ε is a set that μ take maximum range in $B_{\varepsilon(\mu)}$.

REFERENCES

Isaiah Andrews and Anna Mikusheva. GMM is inadmissible under weak identification. *arXiv preprint arXiv:2204.12462*, 2022a.

Isaiah Andrews and Anna Mikusheva. Optimal decision rules for weak gmm. *Econometrica*, 90 (2):715–748, 2022b.

Timothy B Armstrong and Michal Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021.

David H Autor, Susan N Houseman, and Sari Pekkala Kerr. The effect of work first job placements on the distribution of earnings: An instrumental variable quantile regression approach. *Journal of Labor Economics*, 35(1):149–190, 2017.

Andrew Barron, Mark J Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

Daniel Berkowitz, Mehmet Caner, and Ying Fang. The validity of instruments revisited. *Journal of Econometrics*, 166(2):255–266, 2012.

Stéphane Bonhomme and Martin Weidner. Minimizing sensitivity to model misspecification. *Quantitative Economics*, 13(3):907–954, 2022.

Xiaohong Chen, Elie T Tamer, and Alexander Torgovitsky. Sensitivity analysis in semiparametric likelihood models. *Cowles foundation discussion paper*, 2011.

Xiaohong Chen, Timothy M Christensen, and Elie Tamer. Monte carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018, 2018.

Xu Cheng, Zhipeng Liao, and Ruoyao Shi. Uniform asymptotic risk of averaging gmm estimator robust to misspecification, second version. 2015.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.

Victor Chernozhukov and Han Hong. An mcmc approach to classical estimation. *Journal of econometrics*, 115(2):293–346, 2003.

Victor Chernozhukov, Han Hong, and Elie Tamer. Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284, 2007.

Timothy G Conley, Christian B Hansen, and Peter E Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012.

Dennis D Cox. An analysis of bayesian inference for nonparametric regression. *The Annals of Statistics*, pages 903–923, 1993.

Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

Kjell A Doksum and Albert Y Lo. Consistent and robust bayes procedures for location based on partial information. *The Annals of Statistics*, pages 443–453, 1990.

Jean-Pierre Florens and Anna Simoni. Gaussian processes and bayesian moment estimation. *Journal of Business and Economic Statistics*, 39(2):482–492, 2021.

A. Ronald Gallant. Reflections on the probability space induced by moment conditions with implications for bayesian inference. *Journal of Financial Econometrics*, 14(2):229–247, 2016.

Edward L Glaeser, Sari Pekkala Kerr, and William R Kerr. Entrepreneurship and urban growth: An empirical assessment with historical mines. *Review of Economics and Statistics*, 97(2):498–520, 2015.

Paul Gustafson. Bayesian inference for partially identified models. *The international journal of biostatistics*, 6(2), 2010.

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

Bruce E Hansen and Seojeong Lee. Inference for iterated gmm under misspecification. *Econometrica*, 89(3):1419–1447, 2021.

Christian Hansen, James B McDonald, and Whitney K Newey. Instrumental variables estimation with flexible distributions. *Journal of Business & Economic Statistics*, 28(1):13–25, 2010.

Lars Peter Hansen and Thomas J Sargent. Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics*, 4(3):519–535, 2001.

Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton university press, 2008.

Lars Peter Hansen and Thomas J Sargent. Fragile beliefs and the price of uncertainty. *Quantitative Economics*, 1(1):129–162, 2010.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Jae-Young Kim. Limited information likelihood and bayesian analysis. *Journal of Econometrics*, 107:175–193, 2002.

Matthew A Masten and Alexandre Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020.

Hyungsik Roger Moon and Frank Schorfheide. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012.

Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

Vladimir Spokoiny. Penalized maximum likelihood estimation and effective dimension. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):389–429, 2017.