cemmap
centre for microdata methods and practice

Plausible GMM: a quasi-bayesian approach

Victor Chernozhukov
Christian B. Hansen
Lingwei Kong
Weining Wang

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP14/25

Economic and Social Research Council

PLAUSIBLE GMM: A QUASI-BAYESIAN APPROACH

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 framework where researchers model their belief about the potential degree of misspecification via a prior distribution and adopt a quasi-Bayesian approach for performing inference on structural parameters. We provide quasi-posterior concentration results, verify that quasi-posteriors can be used to obtain approximately optimal Bayesian decision rules under the maintained prior structure over misspecification, and provide a form of frequentist coverage results. We illustrate the approach through empirical examples where we obtain informative inference for structural objects allowing for substantial relaxations of the requirement that moment conditions hold exactly. **Keywords**: sensitivity analysis, misspecification, generalized method of moment (GMM),

quasi-Bayes, Bernstein–von Mises theorem (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 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 whether there are unmodeled sources of confounding or sources of misspecification that would

---

Date: June 30, 2025.
[a]: Department of Economics, Massachusetts Institute of Technology.
[b]: Booth School of Business, The University of Chicago.
[c]: Department of Economics, Econometrics and Finance, University of Groningen.
[d]: Department of Economics, University of Bristol.
We thank Bertille Antonie, Tom Boot, Peter Boswijk, Songxi Chen, Aureo de-Paula, Artūras Juodis, Byunghoon Kang, Frank Kleibergen, Michal Kolesár, Jun Liu, Anna Mikusheva, Whitney K. Newey, Frank A. G. Windmeijer, Taisuke Otsu, Johannes Schmidt-Hieber, and Liyang Sun for thoughtful comments. All errors remain on our own.

result in moment conditions failing to hold exactly in any given empirical setting. Unfortunately, it is 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, 2008, 2010), Hall and Inoue (2003), Cheng et al. (2019), and Hansen and Lee (2021).

In this paper, we consider one approach to estimation and inference within a semiparametric structural model characterized by a set of moment restrictions, allowing for the possibility that the specified moment conditions do not hold exactly. We consider a setting where a researcher has access to an observable independent and identically distributed (i.i.d.) data stream $\{Z_t\}_{t=1}^T$ realized from unknown distribution $\mathbb{P}_{\mu_*}$.[1] The researcher specifies a structural model, defined in terms of an economically meaningful k-dimensional parameter vector θ_*, that restricts the joint distribution via moment conditions $m(\theta_*) = \mathbb{E}_{\mathbb{P}_{\mu_*}}[g(Z_t,\theta_*)] = \mu_*$ for some $q \geq k$ dimensional vector μ_*. Of course, informative inference about θ_* is impossible without beliefs about μ_*. Classical estimation and inference proceed under the dogmatic prior belief that $\mu_* \equiv 0$.

Rather than adopt dogmatic prior beliefs, we conceptualize the notion that moment restrictions are plausible – but not known to hold exactly – by assuming the researcher is able to place an informative, but not dogmatic, prior distribution over μ_*. The use of a proper prior over μ_* allows informative inference about model parameters to proceed while relaxing the usual restriction that $\mu_* \equiv 0$. By concentrating this prior over 0, we capture the notion that a researcher subjectively believes the structural moment restrictions are “likely” to be correct. The spread and shape of the prior away from 0 further captures the researchers’ beliefs about “likely” economically motivated possible deviations from the baseline structural model. Thus, the use of a proper prior distribution over μ_* provides a way for researchers to explicitly encode their subjective beliefs over the plausibility of their structural model.

Given that we wish to only leverage structural moment conditions and choose to conceptualize the plausibility of these moment conditions by using a proper prior distribution, it is natural to consider estimation and inference based on approximate or quasi-Bayesian posteriors (QBP), as in Kim (2002) and Chernozhukov and Hong (2003).[2] QBPs provide a tractable approach to approximate Bayesian estimation and inference in traditional semiparametric moment condition models where $\mu_* \equiv 0$ is imposed; see, e.g., Kim (2002), Gallant (2016), Florens and Simoni (2021), and Andrews and Mikusheva (2022). Outside of the Bayesian motivation, Chernozhukov and Hong (2003) demonstrate that these estimators have desirable frequentist properties within the moment condition framework when $\mu_* \equiv 0$. In addition, Andrews and Mikusheva (2022) verifies that quasi-Bayes decision rules approximate Bayes’ optimal decision rules within the weakly identified GMM framework.

[1]While we maintain the assumption that the Z_t are independent and identically distributed, some of our theoretical results hold more generally. We briefly comment on extensions to non-i.i.d. settings in Section 4.3.

[2]In Chernozhukov and Hong (2003), estimators produced from quasi-Bayesian posteriors are referred to as Laplace-Type Estimators (LTEs). These approaches are also connected to "probably approximate correct inference", e.g., Catoni (2004).

In this paper, we extend the QBP framework to deal with settings where moment conditions are not assumed to hold exactly, i.e., to settings with non-dogmatic prior over μ_*. We refer to estimation and inference within this setting as plausible GMM (PGMM). A central challenge in this setting is that θ_* and μ_* are not jointly identified, which implies that the impact of priors is not asymptotically negligible. We develop new technical results that address this challenge and show that, under suitable regularity conditions, key properties of the QBP framework extend to the PGMM setting. Building on Andrews and Mikusheva (2022), we verify that quasi-Bayes decision rules approximate Bayes optimal decision rules given the provided subjective priors. We also generalize the results of Chernozhukov and Hong (2003) to verify that interval estimates from QBPs have a well-defined *ex-ante* notion of frequentist coverage under a sampling process where nature first draws a degree of misspecification from the subjective prior for μ_* and then the model realizes conditional on this draw as in Conley et al. (2012) and analogous to the coverage notion considered in Andrews et al. (2024). Finally, we provide novel large sample approximations for quasi-posterior distributions within this partially identified framework, allowing for the dimension of both θ_* and μ_* to increase with sample size. These results can be viewed as new Bernstein-von Mises type theorems that explicitly account for additional terms that arise when dealing with misspecification.

We illustrate the use of QBPs with proper priors over the degree of misspecification, μ_*, via two empirical examples. 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. The empirical applications demonstrate that one can still draw economically meaningful conclusions using our approach in real applications under what we consider to be sensible beliefs about model misspecification. The approach thus potentially enhances the credibility of the qualitative empirical results. We also use the empirical examples to discuss prior choice, illustrate the impact of prior choice on the resulting quasi-posteriors, and discuss empirically motivated sensitivity analysis.

There is a large existing literature on sensitivity analysis and partial identification. Much of this research focuses on establishing formal frequentist guarantees for estimating and performing inference on the identified set. See, e.g., Canay and Shaikh (2017) and Molinari (2020) for excellent reviews and Norets and Tang (2014), Kline and Tamer (2016), Chib et al. (2018), Liao and Simoni (2019), Giacomini and Kitagawa (2021), Giacomini et al. (2022), and Kuang (2024) for approaches leveraging Bayesian methods.

Within this literature, our work is closely related to Armstrong and Kolesár (2021). Armstrong and Kolesár (2021) also considers a moment condition model where, under correct specification, $m(\theta_*) = 0$ for θ_* the true population parameter value in their setup. They then allow for misspecification by allowing for the possibility that $m(\theta_*) = C_T \neq 0$ where the exact value of C_T is unknown but taken to be an element of a known set; see also Bonhomme and Weidner (2022). Armstrong and Kolesár (2021) focuses on the setting where misspecification is taken to be of the similar magnitude as sampling error, where $C_T = c/\sqrt{T}$, though their approach also applies when misspecification is not modeled as depending on the sample size. Armstrong and Kolesár

(2021) provides a tractable approach to obtain valid frequentist inference, discusses performing sensitivity analysis by varying the set to which C_T is assumed to belong, and shows that optimal GMM weighting matrices in this setting should trade off sampling uncertainty in the moments with the magnitude of potential misspecification. In an analogous result, we verify that the QBPs center on a GMM estimator that uses a weighting matrix that trades off moment precision with misspecification as in Armstrong and Kolesár (2021) in a setting with Gaussian priors with prior variance proportional to $1/T$, though the exact structure differs.

Another related paper is Chen et al. (2018), which uses simulation from quasi-posteriors to develop confidence sets that have frequentist coverage guarantees for identified sets in general settings that include moment condition models. Chen et al. (2018) illustrate how to use their approach in moment inequality models by augmenting the model with auxiliary parameter μ_*. Rather than impose a proper prior over μ_*, Chen et al. (2018) use the known support restrictions from the moment inequalities and profile out μ_*. Their procedure could readily be adapted to settings with other support restrictions on μ_*, including the setting of our paper if valid frequentist inference for the identified set under support restrictions is the ultimate goal. To establish their results, Chen et al. (2018) develop Bernstein-von Mises type theorems for quasi-posteriors under partial identification. We complement this contribution by establishing similar results in the setting with a proper prior over μ_*. As the proper prior over μ_* has important impacts on posterior concentration, our formal results use different theoretical development relative to the formal results in Chen et al. (2018) which may be of independent interest.

Our perspective is different from much of this previous work whose chief goal is establishing frequentist guarantees under partial identification as we wish to impose a proper subjective prior over μ_*. That is, we mostly maintain a subjective Bayesian perspective as we believe there are settings where researchers will want to employ informative, subjective beliefs about potential misspecification. Our work thus aligns closely with the strand of Bayesian work on partial identification reviewed in Gustafson (2015). An element of this work is establishing posterior concentration results. We contribute to this literature by providing such concentration results within the semiparametric moment condition framework where the source of partial identification is uncertainty about the potential misspecification. These concentration results also allow us to consider frequentist properties of posterior summaries and thus complement the broader literature on partial identification and sensitivity analysis.

Within the Bayesian literature on partial identification and misspecification, our setup resembles Chib et al. (2018). Chib et al. (2018) consider a Bayesian semiparametric moment condition model that includes an auxiliary parameter equivalent to μ_* to capture the misspecification of some moment conditions. However, Chib et al. (2018) focus on establishing posterior concentration on a well-defined pseudo-true value in the case of model misspecification, which requires that the number of free elements in μ_* is less than $q-k$. We instead allow for the possibility that all elements of μ_* are free, which precludes point identification of even a pseudo-true value and complicates establishing asymptotic concentration. Our formal results differ substantially in that priors have a non-negligible impact on asymptotic results, and posteriors do not generally concentrate on a unique pseudo-true value.

Finally, in interesting recent work, Andrews et al. (2024) consider Bayesian decision making in a population minimum distance problem under potential misspecification. Andrews et al. (2024) focus on the case where $q > k$ and, among other results, provide a class of priors and a confidence interval construction based on enlarging standard intervals in a way that depends on the standard minimum distance overidentification test statistic that provides *ex-ante* coverage under these priors. Similar to Andrews et al. (2024), we also establish a sense of *ex-ante* coverage, though our results hold within the semiparametric moment condition framework under our subjective prior over the degree of misspecification and cover the case where $q = k$.

The remainder of the paper is organized as follows. In Section 2, we more carefully discuss the main ideas and provide a convenient approximation result for the case when the prior for μ_* is taken to be Gaussian with precision proportional to the sample size – the case of "local misspecification." We present the empirical illustrations in Section 3, and we provide formal results in Section 4. We present further details and proofs in the Appendix and the Supplemental Appendix.

2. The Approach: Main Ideas

2.1. **Plausible Moment Restriction Model.** Suppose that we observe data $\{Z_t\}_{t=1}^T$ which are a realization from some unknown distribution $\mathbb{P}_{\mu_*}$. Suppose that we also have a posited structural economic model, which provides a set of moment restrictions on the distribution $\mathbb{P}_{\mu_*}$ indexed by a q-dimensional parameter $\mu_* \in \mathcal{M}$. Specifically, suppose the structural model implies a set of $q \geq k$ equations for a k-dimensional parameter $\theta \in \Theta$

$$m(\theta) = \mathbb{E}_{\mathbb{P}_{\mu_*}}[g(Z_t, \theta)],$$

such that there exists a vector μ_* corresponding to a target parameter θ_* satisfying

$$m(\theta_*) = \mu_*.$$

Of course, with no restrictions on the vector μ_*, it is impossible to update beliefs about θ_* or the distribution $\mathbb{P}_{\mu_*}$ using the structural model. For any posited value of θ and distribution $\mathbb{P}_\mu$, we can always set $\mu = \mathbb{E}_{\mathbb{P}_\mu}[g(Z_t, \theta)]$ such that the structural moment equation is satisfied.[3] Classical approaches to moment restriction models bypass this difficulty by assuming the vector μ_* is known to be a fixed, prespecified vector (without loss of generality $\mu_* \equiv 0$). This classical approach is equivalent to imposing the dogmatic prior that the researcher knows the structural moment equations hold exactly under $\mathbb{P}_{\mu_*}$ – that is, the researcher has a dogmatic prior that the moment equations are correctly specified.

[3] Priors over θ_* and $\mathbb{P}_{\mu_*}$ produce restrictions over μ_*, but the structural moment restriction adds no additional information if μ_* is left completely unrestricted.

Unfortunately, it is hard to be fully confident that a set of structural moment restrictions hold exactly in many settings. For example, we may worry that there are unobserved confounding variables or that the functional form of the structural model is incorrect. We allow for departures from the dogmatic belief that the structural moment restrictions hold exactly by making use of a proper, non-degenerate prior distribution over μ_*, denoted $\pi(\mu)$. The use of a proper prior over μ_* allows moment restrictions to be informative in updating beliefs about θ_* while falling short of imposing the often implausible restriction that moment restrictions hold exactly.

As a concrete example, consider the constant coefficient linear model

$$Y_t = X_t\theta_* + U_t,$$

where X_t is an observed variable with $\mathbb{E}_{\mathbb{P}_{\mu_*}}[X_t U_t] \neq 0$. Further, suppose we observe an additional variable D_t that, based on economic reasoning or institutional knowledge, we believe satisfies the usual instrument exclusion restriction $\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t U_t] = 0$ for $t = 1, ..., T$. Under this belief, we obtain the moment restriction $\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t U_t] = 0$ which can be used to identify the structural parameter θ_*.

However, it is hard to know that the IV exclusion restriction holds exactly in many settings. For example, we might worry that there exists an unobserved confound, M_t, that covaries with both Y_t and D_t such that $U_t = M_t + V_t$, $\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t M_t] = \mu_* \neq 0$ and $\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t V_t] = 0$. Imposing the moment restriction $\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t U_t] = \mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t(Y_t - \theta X_t)] = 0$ and solving for θ produces

$$\theta = \left(\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t X_t]\right)^{-1} \mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t Y_t] = \theta_* + \left(\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t X_t]\right)^{-1} \mu_* \neq \theta_*.$$

Within the IV example, we might instead consider the restriction $\mathbb{E}_{\mathbb{P}_{\mu_*}}[D_t(Y_t - \theta_* X_t)] = \mu_*$ where we assume that μ_* is a fixed realization from a random variable μ, e.g., $\mu \sim N(0, \sigma^2)$. Here, the assumed distribution captures the notion that the researcher believes the instrument is "close to" being valid in that the prior mass for μ is concentrated around 0. The distribution also encapsulates that the researcher believes it is incredibly unlikely that the moment restriction is perfect as $\{\mu = 0\}$ occurs with zero probability under such a distribution. Finally, the researcher can control beliefs about the strength of the unobserved confounder via the prior variance, σ^2, while technically allowing for μ_* to be unbounded. That is, the proper prior over μ_* allows a well-defined and concrete description of the moment restriction being plausibly, but not certainly, satisfied.

To summarize, we are interested in a formalized version of a "plausible" moment restriction model characterized by parameters (θ, μ) such that

$$m(\theta) = \mu$$

and μ is governed by a prior distribution with density $\pi(\mu)$. We refer to μ as the "plausibility characteristic," and we denote any root of the equation $m(\theta) = \mu$ as $\theta(\mu)$. For establishing some of the formal results in Section 4, we will assume that $\pi(\mu)$ places strictly positive mass over a region Γ such that solutions $\theta(\mu)$ exist for $\mu \in \Gamma$. This prior restriction is essentially trivially satisfied for any prior when $q = k$; see, e.g. Hall and Inoue (2003). However, satisfaction of this assumption is not guaranteed with $q > k$, suggesting that care should be taken in adding moment conditions about which a researcher has relatively weak prior beliefs unless the researcher is willing to use very diffuse priors.[4]

In the next section, we outline a quasi-Bayesian approach to perform inference on our main target parameter θ.

2.2. **(Quasi-)Bayes for Plausible Moment Restrictions.**

We adopt a (quasi-)Bayesian approach to performing inference within the plausible moment restriction model. Let

$$\widehat{m}(\theta) := \frac{1}{T}\sum_{t=1}^{T} g(Z_t, \theta)$$

be the average of $g(Z_t,\theta)$ against the empirical distribution at a given value of θ. We can then define a continuous updating GMM-type criterion function for parameters (θ,μ) as

$$Q_T(\theta,\mu) = -T\left(\widehat{m}(\theta)-\mu\right)^\top \widehat{\Omega}_T(\theta)^{-1}\left(\widehat{m}(\theta)-\mu\right) \tag{1}$$

for $\widehat{\Omega}_T(\theta)$ a positive definite matrix approximating

$$\Omega(\theta) = \lim_{T\to\infty} \mathrm{Var}(\sqrt{T}(\widehat{m}(\theta) - m(\theta))).$$

For example, it would make sense to use

$$\widehat{\Omega}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(g(Z_t,\theta)-\widehat{m}(\theta)\right)\left(g(Z_t,\theta)-\widehat{m}(\theta)\right)^\top$$

under the assumption that the Z_t are i.i.d.

A quasi-posterior based on criterion (1) is then obtained as

$$p_T(\theta,\mu) = \frac{\exp\left(\frac{1}{2}Q_T(\theta,\mu)\right)\pi(\theta,\mu)}{\int_\Xi \exp\left(\frac{1}{2}Q_T(\theta,\mu)\right)\pi(\theta,\mu)\,d\mu\, d\theta} \tag{2}$$

where $\pi(\theta,\mu) = \pi(\theta|\mu)\pi(\mu)$ is the joint prior over (θ,μ) and Ξ is the corresponding (joint) prior support. We expect researchers will often specify constant priors over θ – i.e. set $\pi(\theta|\mu) \propto 1$. However, just as it does with μ, the quasi-Bayes approach offers a convenient avenue for imposing economically motivated priors over θ, which may be desirable in some settings. In general, $p_T(\theta,\mu)$ will not be available analytically but will need to be approximated using Markov Chain Monte Carlo (MCMC) or other sampling methods; see, e.g., Robert and Casella (2005) for a classic textbook introduction. We also provide a simple Gaussian approximation to the posterior in a setting where the prior for μ is taken to be normal with a small variance in Section 2.3.

[4]From a frequentist perspective, Armstrong and Kolesár (2021) note that usual overidentification statistics can be used to infer lower bounds on the magnitude of μ. Andrews et al. (2024) consider an interesting different approach in overidentified settings that operationally inflates the size of confidence sets based on the magnitude of overidentification statistics.

We obtain marginal posteriors for θ and μ as usual by integration:

$$p_T(\theta) = \int_{\mathscr{M}} p_T(\theta,\mu)d\mu \text{ and } p_T(\mu) = \int_{\Theta} p_T(\theta,\mu)d\theta,$$

where $\mathscr{M}$ and Θ respectively denote the support of μ and θ. $p_T(\theta)$ captures posterior information about the economically meaningful parameter θ and thus is the chief object of interest. $p_T(\mu)$ also potentially provides useful information in summarizing posterior beliefs about the plausibility term μ.[5]

Given the quasi-posterior, we can also immediately formulate optimal decisions under the quasi-posterior by minimizing the quasi-posterior expected risk. Specifically, let $\ell(\theta,\mu,d)$ be a loss function that depends on the underlying model parameters (θ,μ) and a decision $d \in \mathscr{D}$.[6] We can then define the expected loss minimizing decision under the quasi-posterior, denoted $s_T(p_T)$, as

$$s_T(p_T) \in \arg\min_{d\in\mathscr{D}} \int_{\Xi} \ell(\theta,\mu,d)\, p_T(\theta,\mu)\, d\mu d\theta. \tag{3}$$

The quasi-posterior (2) and inferential objects obtained from it, such as optimal decisions or credible intervals, can be given an approximate Bayesian interpretation. Florens and Simoni (2021) and Andrews and Mikusheva (2022) study the classic semiparametric moment condition model under correct specification such that $m(\theta_*) \equiv 0$, i.e., under the dogmatic prior that $\mu \equiv 0$. Florens and Simoni (2021) provide prior choices for the unknown model density such that the analog of (2) within this setting corresponds to the posterior for θ as the limit when the prior becomes diffuse. By augmenting the parameter space to include μ, (2) can be obtained as the posterior over (θ,μ) under the prior structure of Florens and Simoni (2021). Andrews and Mikusheva (2022) study optimal decision rules in weakly identified moment condition models under correct specification such that $m(\theta_*) \equiv 0$. Andrews and Mikusheva (2022) establish that the analog of (2) for their setting, corresponding to (2) under the dogmatic prior that $\mu \equiv 0$, results as the limit of a sequence of posteriors under a specific choice of proper priors. The resulting quasi-Bayes decision rule then corresponds to the pointwise limit of the sequence of Bayes decision rules and can thus be motivated as approximating the optimal Bayes decision. We show that these results continue to apply in our setting with non-dogmatic prior over μ in Section 4. Of course, given the non-degenerate prior over μ, the optimal Bayes decision depends explicitly on not only the prior for θ as in Andrews and Mikusheva (2022) but also on the prior for μ. See also Kim (2002) and Gallant (2016) for additional Bayesian motivation and perspective.

From a purely frequentist perspective, Chernozhukov and Hong (2003) verify that inference based on the quasi-posterior is asymptotically equivalent to inference based on efficient GMM in strongly identified settings under correct specification ($m(\theta_*) \equiv 0$). Chernozhukov and Hong (2003) further argue that basing frequentist estimation and inference on posterior summaries

[5]Note that, while θ and μ are not jointly identified, the imposition of a proper prior over either θ or μ will lead to posterior updating over both θ and μ, even in settings with $q = k$.

[6]In most applications, this loss function will depend only on the economically motivated parameter θ, but we allow the loss to be over μ as well.

from (2), such as using the posterior mean as a point estimator and posterior credibility interval as a confidence interval, may be desirable in settings where (1) is hard to optimize. However, credible intervals resulting from the quasi-posterior (2) within the partially identified setting where μ follows a non-degenerate prior no longer generally deliver usual frequentist coverage guarantees, though they do still have an approximate Bayes interpretation; see Moon and Schorfheide (2012) and Gustafson (2015).[7]

We extend the results from Chernozhukov and Hong (2003) by studying the frequentist properties of the quasi-Bayes posterior within our setting with a non-dogmatic prior over μ allowing for sequences with increasing k and q in Section 4. Within this setting, we provide new Bernstein-von Mises type posterior concentration results showing that the quasi-posterior converges to a mixture of Gaussian distributions where the mixture weights and components depend heavily on the specific prior $\pi(\mu)$. That is, the posterior aligns with our intuition about partial identification in that the prior for μ plays a key role in the shape of the posterior even in the limit. See also Gustafson (2015). In the case that one has a dogmatic prior over μ—e.g., $\mu \equiv \mu_*$—our concentration result reproduces Chernozhukov and Hong (2003), in the sense that our posterior approximation collapses to a Gaussian random variable with center $\theta(\mu_*)$ and variance equal to the limiting variance of the efficient GMM estimator in this case.

Based on the posterior concentration results, we then have that usual Bayesian credible regions from the quasi-posterior (2) have correct frequentist coverage within a two-stage sampling thought experiment where each repeated sample corresponds to drawing a value μ_* from a random variable with density $\pi(\mu)$ and then generating data such that $m(\theta(\mu_*)) = \mu_*$. Alternatively, one can view this notion of coverage as providing an *ex ante* coverage guarantee in a setting where a single value μ_* will be realized from a random variable with density $\pi(\mu)$.

Given that the two-stage sampling notion of coverage is non-standard, we consider a final approach to leveraging (2) to provide a confidence set with a uniform frequentist coverage guarantee when the plausibility characteristic is taken to be some fixed vector, μ_0, whose value is unknown, but where it is known that $\mu_0 \in C$ for some known compact set C. The basic idea is that we can use QBPs exactly as in Chernozhukov and Hong (2003) to obtain point and interval estimates that would be asymptotically equivalent to efficient GMM for any fixed $\mu_* \in C$ under strong identification. Letting $CI(\mu_*, \alpha)$ be the resulting $(1-\alpha)\%$ credible interval, it then follows that $\cup_{\mu_* \in C} CI(\mu_*, \alpha)$ has coverage at least $(1-\alpha)\%$ for $\theta(\mu_0)$. This approach mimics, e.g., the union of confidence intervals approach from Conley et al. (2012) and the approach outlined in Remark 3.3 of Armstrong and Kolesár (2021).

[7]Andrews and Mikusheva (2022) also show that quasi-posterior interval estimates generally do not provide correct frequentist coverage in weakly identified settings and suggest a procedure to obtain confidence sets with proper frequentist coverage. We choose to focus our coverage results on strongly identified settings while allowing for a non-dogmatic prior over μ. In principle, the weak identification robust confidence set construction of Andrews and Mikusheva (2022) could also be incorporated in the present setting.

2.3. **Simple Quasi-Bayes Inference using Gaussian Local Priors.** In this section, we outline an approach to doing approximate quasi-Bayes inference when the prior for the plausibility term μ is normal with a small variance. Specifically, suppose that our prior is

$$\mu \sim \mathcal{N}\left(\mu_0, \frac{\Lambda}{T}\right) \tag{4}$$

for some fixed q dimensional vector μ_0 and a fixed, full-rank $q \times q$ matrix Λ.[8] A simple form of Λ is a diagonal matrix with λ_k's on the diagonal, where a small λ_k indicates that there is little uncertainty about the plausibility of the k-th moment and larger values indicate higher uncertainty.

Intuitively, prior (4) captures the case where misspecification is believed to be small but non-zero in the sense that we believe the moment conditions "almost" hold with $m(\theta_0) = \mu_0$. Considering a sequence of priors with variance of order $1/T$ means that prior uncertainty concentrates at the same rate as the sample moments, so neither will dominate as we consider large T asymptotic approximations. Following the literature, we refer to (4) as a local prior; see, e.g. Conley et al. (2012) and Armstrong and Kolesár (2021).

We now provide an approximation to the quasi-posterior with $\pi(\mu)$ defined in (4) and a flat prior over θ. We assume μ_0 is such that a solution $\theta(\mu_0)$ satisfying $m(\theta(\mu_0)) = \mu_0$ exists. For simplicity, we further set $\mu_0 = 0$. We assume strong identification in the sense that $m(\theta(\mu)) = \mu$ has unique solution $\theta(\mu)$ for each μ and that the following linearization around $\theta_0 \equiv \theta(\mu_0)$ holds:

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

where $G = \partial\mathbb{E}(\widehat{m}(\theta))/\partial\theta|_{\theta=\theta_0}$ and $G^\top G$ has minimal eigenvalue bounded away from zero. Further, define the weighting matrix $\widehat{A}_{T,\theta}$ and its population counterpart A_θ:

$$\begin{aligned}
\widehat{A}_{T,\theta} &= \widehat{\Omega}_T(\theta)^{-1} - \widehat{\Omega}_T(\theta)^{-1}[\Lambda^{-1} + \widehat{\Omega}_T(\theta)^{-1}]^{-1}\widehat{\Omega}_T(\theta)^{-1}, \\
A_\theta &= \Omega(\theta)^{-1} - \Omega(\theta)^{-1}[\Lambda^{-1} + \Omega(\theta)^{-1}]^{-1}\Omega(\theta)^{-1}.
\end{aligned}$$

Under these conditions, we show in Section 4 that $p_T(\theta)$ is approximately proportional to,

$$\exp\left(-T\|\widehat{m}(\widehat{\theta}) + G(\theta - \widehat{\theta})\|^2_{\widehat{A}_{T,\theta}}/2\right),$$

where the quasi-posterior mode $\widehat{\theta}$ is the GMM estimator obtained using weighting matrix $\widehat{A}_{T,\theta}$. That is, we can approximate the quasi-posterior for θ as

$$\theta \approx \mathcal{N}\left(\widehat{\theta}, \frac{V}{T}\right), \quad V = \left(G^\top \widehat{A}_{T,\theta} G\right)^{-1}. \tag{5}$$

Note that the weighting matrix A_{θ_0} in the quasi-posterior is different from the standard efficient GMM weighting matrix $\Omega(\theta_0)^{-1}$.[9] Specifically, A_{θ_0} reflects additional uncertainty brought by not having a fixed, known value for μ. We do see that we recover the case of efficient GMM by letting $\Lambda \to 0$—i.e. by assuming that there is no uncertainty over the moment conditions.

---

[8]We could relax the restriction that Λ is full-rank at the cost of a modest complication of notation.

[9]As $\widehat{A}_{T,\theta}$ will converge to A_{θ_0} with $T \to \infty$, we provide intuition as if A_{θ_0} were known.

The quasi-posterior has several interesting features. The center of the quasi-posterior, $\widehat{\theta}$, corresponds to the classical GMM estimator that uses the weighting matrix A_{θ_0} rather than the efficient weighting matrix $\Omega(\theta_0)^{-1}$. This centering is intuitive as $\Omega(\theta_0)$ captures only sampling variation in the moments but does not reflect the additional uncertainty arising from the plausibility of the moments. The weighting matrix A_{θ_0} incorporates both sources of uncertainty, intuitively placing the most weight on moments about which the researcher is most confident in the sense that combined sampling variability and plausibility uncertainty is lowest.

Looking at the quasi-posterior variance, there are two further noteworthy features. First, the variance matrix $V = (G^\top A_{\theta_0} G)^{-1} \geq (G^\top \Omega(\theta_0)^{-1} G)^{-1}$, where $(G^\top \Omega(\theta_0)^{-1} G)^{-1}$ is the usual asymptotic variance of the efficient GMM estimator. The variance matrix V thus captures additional uncertainty, relative to efficient GMM, introduced by a lack of certainty over the validity of the moment restrictions. Second, the approximate sampling distribution of $\widehat{\theta}$ is

$$\sqrt{T}(\widehat{\theta} - \theta_0) \to_d \mathcal{N}(0, \bar{V}), \quad \bar{V} = (G^\top A_{\theta_0} G)^{-1} G^\top A_{\theta_0} \Omega(\theta_0) A_{\theta_0} G (G^\top A_{\theta_0} G)^{-1},$$

where $\bar{V} \leq V$ because $A_{\theta_0}\Omega(\theta_0)A_{\theta_0} \leq A_{\theta_0}$. Thus, the quasi-posterior variance is also larger than the asymptotic variance of the A_{θ_0}-weighted GMM estimator. Again, this larger quasi-posterior variance arises because the sampling distribution of the A_{θ_0}-weighted GMM estimator is obtained under the dogmatic belief that $\mu \equiv 0$ and thus does not reflect uncertainty about the validity of the moment restrictions outside of through reweighting the moment conditions.

To summarize, the approximation in (5) provides a very simple avenue to obtain approximate Bayesian inference under local Gaussian priors. While restrictive, it does seem like a Gaussian prior with small variance may provide a reasonable model for subjective beliefs about moment condition violations in some settings, and we illustrate the use of the approximation, along with illustrating simulation of the full quasi-posterior, in the empirical examples in the next section. More importantly, the approximation captures the clear intuition that there is no free lunch. Incorporating non-dogmatic priors over moment condition violations naturally results in less informative inference relative to the case where dogmatic priors are imposed – reflecting the researcher's uncertainty about the validity of the moment restrictions. This property seems desirable as the resulting inference likely more accurately reflects what can be learned in real empirical settings where model uncertainty exists.

3. Empirical Applications

This section applies plausible GMM in two illustrative empirical applications. In the first, we revisit Acemoglu et al. (2001), which uses linear instrumental variables (IV) to study the effect of institutions on economic output in a relatively small sample. In the second, we revisit Chernozhukov and Hansen (2004), which uses IV quantile regression to examine the effects of 401(k) participation on measures of household assets.

3.1. **Linear IV Example: Effect of Institutions on GDP.** We start by illustrating our methodology by revisiting the classic study of Acemoglu et al. (2001), which investigates the effect of institutions on economic performance.The outcome variable Y_t is the log of PPP-adjusted GDP per capita in 1995 where $t = 1,\ldots,64$ indexes a set of countries that are ex-European colonies. The main regressor of interest, X_t, is a ten-point index measuring protection against expropriation risk, serving as a proxy for institutional quality. We consider a baseline specification from Acemoglu et al. (2001) which includes normalized distance from the equator, W_t, as a control for geographic factors.

To address concerns about the endogeneity of X_t, Acemoglu et al. (2001) adopt an IV strategy. Following Acemoglu et al. (2001), we consider two IV specifications. The first, denoted Linear IV(1), uses the log of settler mortality as the sole instrument. This specification corresponds to the baseline in the original study. The second specification adds the proportion of the population of European descent in 1900 as an additional instrument as is done in a robustness exercise in the original paper. This specification allows us to illustrate our procedure in a setting with an overidentifying moment restriction. We refer to this specification as Linear IV(2).

Formally, we consider the linear IV model

$$Y_t = \alpha + \beta_X X_t + \beta_W W_t + U_t,$$

with parameter vector $\theta = (\alpha, \beta_X, \beta_W)^\top$ and moment condition

$$g(Z_t,\theta) = (1, D_t^\top, W_t)^\top \left(Y_t - \alpha - X_t\beta_X - W_t\beta_W\right),$$

where D_t denotes the vector of instruments.

To implement PGMM, we must specify priors for the parameters (θ,μ). We set the prior for θ as $\mathcal{N}(0, \text{diag}(100, 4, 64))$. We set the prior variances for the elements of θ via a loose argument based on economic intuition. For example, we know that the X_t is measured on a 10-point scale with empirical 25th and 75th percentiles equal to 5.6 and 7.8, respectively, and an empirical range of 3.5 to 10. A coefficient of β_X of approximately .5 would thus suggest moving from the 25th to 75th percentiles of X_t is associated with around a one log unit change (around a 170% change) in GDP, which seems economically quite large. We thus feel comfortable placing a relatively low prior probability on β_X having a magnitude larger than 4. We use the same rationale for our choice of the prior over α and β_W.

To specify the prior for μ, we use reasoning based on the IV model. Specifically, we assume that model misspecification arises from the instrument(s) having a direct effect on the outcome. That is, we consider an "augmented" model

$$Y_t = \alpha + \beta_X X_t + \beta_W W_t + \gamma' D_t + U_t,$$

where γ represents the direct effect of the instrument on the outcome. Considering the setting of Linear IV(1), this augmented structure then suggests that we would specify the moment equation as $\mathbb{E}[g(Z_t,\theta)] = \gamma\mathbb{E}[D_t^2] = \mu$ if we knew γ. If we then were willing to maintain the belief that the justification for exclusion of D_t is convincing enough for us to believe that we should center our prior for γ over no direct effect, we would want to specify a prior centered over 0. Given that D_t is log mortality among Europeans several hundred years prior to 1995, one might also be willing to say that there is relatively little scope for the direct effect of D_t to be large. We benchmark our prior using the subjective belief that, with high probability, the elasticity of GDP with respect to settler mortality is no larger than 10%, corresponding to γ no larger than 0.1. We then encode these prior beliefs by specifying the prior for the entry of μ corresponding to the D_t moment as a mean zero Gaussian with standard deviation $0.05\sqrt{\frac{1}{T}\sum_t D_t^2}$. We believe this provides a sensible benchmark in this example, encoding the prior belief that we found the original argument in Acemoglu et al. (2001) compelling in having prior mass tightly concentrated around 0 but allowing for modest probability of relatively large deviations (say direct effects with magnitudes between 0.05 and 0.1) and relatively small probability of larger deviations. We then complete the prior by assuming the other entries of μ behave similarly to the one corresponding to the instrument. Specifically, our baseline (denoted "PGMM-g") uses a Gaussian prior given by

$$\mu \sim \mathcal{N}(0, \Sigma_T \Omega_d \Sigma_T^\top),$$

where $\Sigma_T = T^{-1}\sum_{t=1}^T (1, W_t, D_t)^\top (1, W_t, D_t)$ and $\Omega_d = 0.05^2 I_3$.

For Linear IV(2), we follow a similar approach. Here, the additional instrument is the proportion of the population of European descent. Assuming its direct impact on the outcome is, with high probability, no greater than 0.01 (a semi-elasticity of 1%), we extend the Gaussian prior construction used in Linear IV(1) by setting

$$\Omega_d = \mathrm{diag}(0.05^2 I_3, 0.005^2),$$

where the final diagonal entry corresponds to the new instrument.

Of course, it is important to gauge sensitivity of the posterior to the prior specification. We thus report results using two additional simple prior settings. In the first, we consider a more diffuse prior for μ (denoted "PGMM(d)-g"), given by $\mathcal{N}(0, c\Sigma_T \Omega_d \Sigma_T^\top)$ with $c = 4$. As a second alternative, we also consider a uniform prior for μ (denoted "PGMM-u"), distributed uniformly over the elliptical region $\mathscr{C} = \{\left(\Sigma_T \Omega_d \Sigma_T^\top\right)^{1/2} c : c'c \leq \chi^2_{0.68}(q)\}$, where q is the number of moment conditions, and $\chi^2_{0.68}(q)$ denotes the 0.68 quantile of $\chi^2(q)$. This prior thus allocates all probability mass to the 68% highest density region of the Gaussian prior used in the "PGMM-g" case.

We report the PGMM quasi-posterior obtained under our different priors, along with the quasi-posterior from Chernozhukov and Hong (2003) obtained under the dogmatic prior $\mu \equiv 0$ (labeled "CH"), for β_X in the specification with one excluded instrument in the top panel of Figure 1.[10] We see that, in terms of β_X, the quasi-posteriors are relatively robust to the prior over μ. As anticipated, we see that the quasi-posteriors become somewhat more diffuse as the prior dispersion increases from $\mu \equiv 0$ to PGMM-g to PGMM(d)-g, although the changes in dispersion are relatively small, despite the large increase in prior dispersion for μ across these cases. Unsurprisingly given the design of the uniform prior, we also see that both the

[10] Quasi-posteriors in the Linear IV(2) case, shown in Figure 4 in the Appendix, exhibit similar patterns. We also present posteriors for elements of μ, which roughly align with the corresponding priors, in Figure 5 in the Appendix.

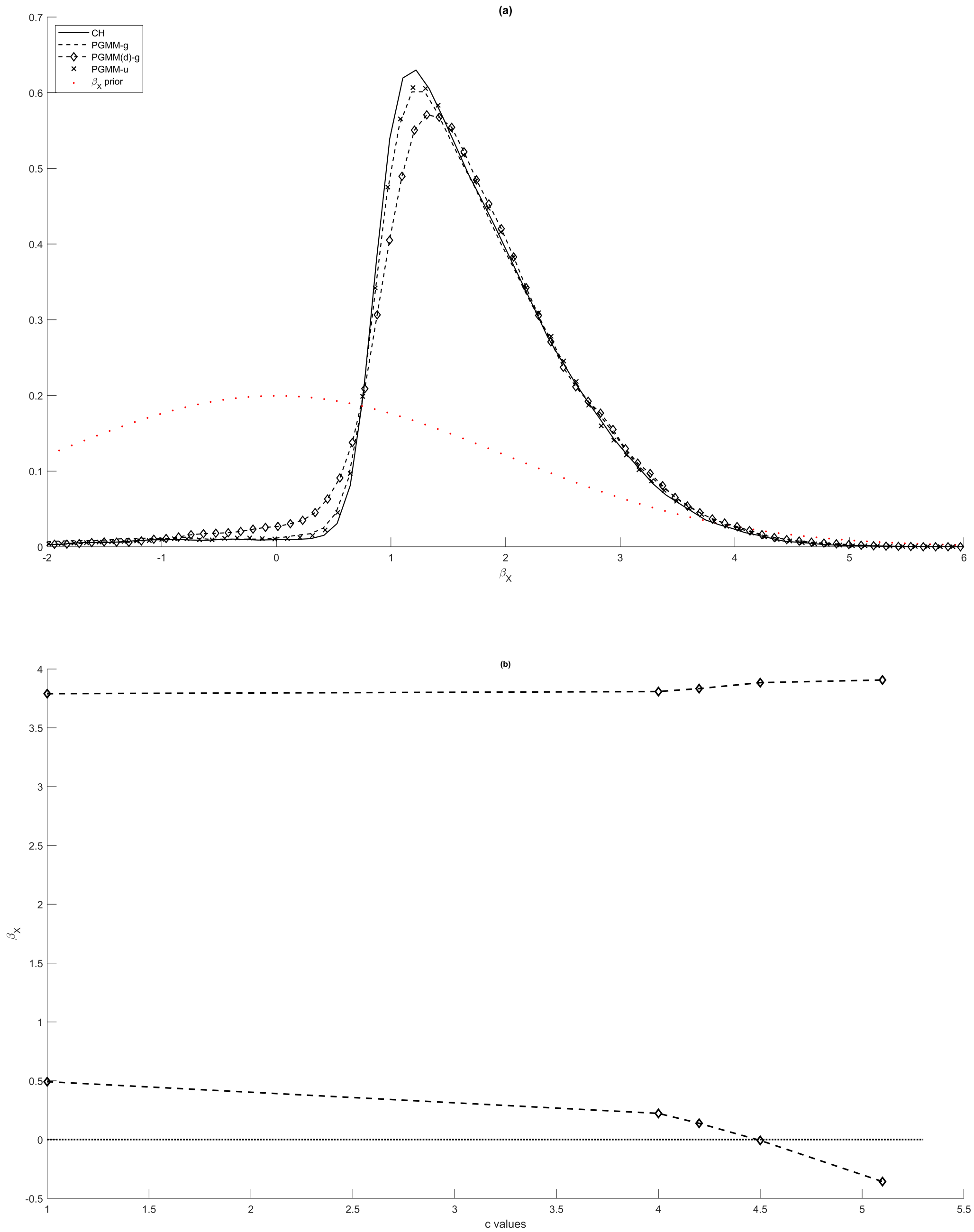

FIGURE 1. Upper panel (a): Linear IV (1) marginal (quasi-)posterior(s) for β_X and its marginal prior (red dotted curve). Lower panel (b): Linear IV(1) 95% HPD intervals for β_X resulting from the PGMM approach along various priors for μ, i.e., $\mathcal{N}(0, c\Sigma_T\Omega_d\Sigma_T^\top)$ for various values of c.

benchmark Gaussian prior (PGMM-g) and related uniform prior (PGMM-u) produce very similar quasi-posteriors for β_X.

For additional insight, we provide 95% HPD intervals for β_X under $\mu \sim \mathcal{N}(0, c\Sigma_T \Omega_d \Sigma_T^\top)$ for additional values of c in the bottom panel of Figure 1. Here, we see that the lower bound of the HPD interval is relatively stable for values of $c \leq 4$. However, the lower bound then decreases relatively quickly as c increases away from 4, crossing 0 at $c \approx 4.5$. That is, under our prior for θ and class of priors for μ, posterior mass remains largely concentrated over positive effects even allowing for what appear to us be economically large deviations from correct specification.

Finally, we observe that the left tails of the quasi-posteriors in Figure 1 are very similar to the upper tail of the maintained prior for β_X. That is, it appears that the behavior of the upper tail of the posteriors may be driven largely by the prior choice for θ. Consequently, the upper bounds of the provided intervals are relatively insensitive to the prior for μ. While unsurprising, we find this interplay between prior structure interesting, especially as researchers often have reasonable economic understanding about plausible values for structural parameters.

We report interval estimates obtained from a variety of procedures under both the Linear IV(1) and Linear IV(2) specification in Table 1. For frequentist methods, we report 95% level confidence intervals, and we report 95% level HPD regions for (quasi-)Bayesian procedures. Specifically, we consider intervals produced by applying the following:

- **2SLS**: two-stage least squares with the usual asymptotic approximation.
- **CUE**: continuous updating estimator with the usual asymptotic approximation.
- **CH**: PGMM under $\mu \equiv 0$; the quasi-Bayes approach from Chernozhukov and Hong (2003).
- **S**: inversion of the S statistic; see Theorem 2 from Stock and Wright (2000).
- **AK**: robust intervals constructed using the local misspecification method of Armstrong and Kolesár (2021) assuming a true value θ_* such that $g(\theta_*) = c/\sqrt{T}, c \in \mathscr{C}$ for $\mathscr{C}$ specified as an in PGMM-u.
- **PGMM-u**: PGMM with uniform prior as previously specified.
- **PGMM-g**: PGMM with baseline Gaussian prior.
- **PGMM(d)-g**: PGMM with diffuse Gaussian prior.
- **Local Approx:** Gaussian limiting approximation for the β_X marginal quasi-posterior of "PGMM-g" under the assumption of local misspecification as described in Equation (5) and Theorem 1.
- **Local Approx (d):** Gaussian limiting approximation for the β_X marginal quasi-posterior of "PGMM(d)-g" under the assumption of local misspecification as described in Equation (5) and Theorem 1.

All approaches allow for heteroskedasticity. 2SLS, CUE, CH, and S maintain the assumption of correct specification. All other procedures allow for departures from correct specification by relaxing the constraint that $\mu \equiv 0$, with AK being frequentist valid and the remaining procedures having a (quasi-)Bayes interpretation. We note that S is formally valid under weak identification, while formal frequentist results for the other procedures are obtained assuming strong identification.

Methods	Linear IV(1)	Linear IV(2)
Assuming correct specification, i.e., $\mu \equiv 0$		
2SLS	[0.56, 1.38]	[0.64, 1.23]
CUE	[0.56, 1.38]	[0.64, 1.21]
CH	[0.61, 3.59]	[0.63, 3.56]
Weak identification robust		
S	[0.63, 3.23]	[0.63, 4.55]
Misspecification robust		
AK	[-28.86, 30.80]	[-0.87, 2.66]
PGMM-u	[0.58, 3.65]	[0.53, 3.78]
PGMM-g	[0.49, 3.79]	[0.54, 3.65]
PGMM(d)-g	[0.22, 3.81]	[0.30, 3.73]
Local Approx	[0.52, 1.42]	[0.53, 1.36]
Local Approx(d)	[0.42, 1.52]	[0.35, 1.57]

TABLE 1. 95% interval estimates for β_X, the effect of institutions on output, obtained from different procedures as described in the main text.

Table 1 shows that, with the exception of the AK interval, the qualitative conclusions are largely consistent across methods: The centers and lower bounds of the intervals lie above zero, suggesting a positive effect of institutions on output. Further, the interval estimates broadly fall within two groups—with 2SLS, CUE, Local Approx, and Local Approx(d) in one and CH, S, PGMM-u, PGMM-g, and PGMM(d)-g in the other. Interestingly, the 2SLS, CUE, and local approximation intervals all rely on asymptotic approximations obtained under strong identification and are substantially narrower than the other intervals. This narrowness may reflect a failure of the conventional asymptotic approximation. In contrast, we obtain the CH and PGMM intervals directly from (quasi-)posteriors without relying on asymptotic approximations. It seems interesting that these intervals are so similar to the interval produced by the weak-identification robust procedure. While this similarity may be coincidental, it is notable and potentially worth further exploration.

Looking at the quasi-Bayes procedures specifically, recall that the CH method imposes the validity of the moment conditions, while the PGMM procedures relax this assumption. This relaxation, of course, results in the PGMM intervals being wider than CH as the PGMM intervals reflect the added uncertainty from accounting for potential misspecification. However, at least under the priors considered, the increase in width is relatively small and does not qualitatively change the conclusions that one would draw relative to CH.

Finally, we observe that the AK intervals lead to qualitatively different conclusions than those from the other approaches. This difference is particularly pronounced in the Linear IV(1) specification, where the AK interval is substantially wider than the intervals produced by the alternative methods. The most informative comparison is between AK and PGMM-u, as both restrict the plausibility term to lie within the same support. The key distinction is that the AK approach is designed to ensure valid frequentist coverage by focusing on least favorable directions within the misspecified model, given only the support restriction on the plausibility term. In contrast, PGMM-u imposes proper subjective priors on both the plausibility term and the structural parameters, θ. In this example, the subjective priors place very little mass on models with economically extreme values of β_X, resulting in the quasi-posterior assigning negligible mass to much of the AK interval. This outcome illustrates how subjective priors can substantially shape the quasi-posterior in partially identified settings. Rather than targeting worst-case combinations, the quasi-posteriors reflect economically motivated beliefs about the joint distribution of the structural parameters and the plausibility term. Both approaches serve meaningful purposes, but we believe there are scenarios in which inference based on the quasi-posterior under subjective priors may offer more economically relevant insights.

3.2. **IV Quantile Regression Example: Effect 401(k) Participation on Financial Assets.** In this subsection, we illustrate the use of PGMM in a non-linear model by using IV quantile regression (IVQR) to estimate the impact of 401(k) participation on quantiles of net financial assets as in Chernozhukov and Hansen (2004). Specifically, we apply PGMM using the IVQR moment conditions from Chernozhukov and Hansen (2005):

$$g_\tau(Z_t, \theta_\tau) = (1, D_t, W_t^\top)^\top \left(\tau - \mathbf{1}\left(Y_t \leqslant \alpha_\tau + X_t \beta_{X,\tau} + W_t^\top \beta_{W,\tau}\right)\right)$$

where τ denotes the quantile of interest; Y_t is the outcome variable representing net total financial assets (in 1991 dollars); X_t is a binary indicator for 401(k) participation; W_t is a vector of control variables; and D_t is a binary instrument indicating 401(k) eligibility.[11] We report results for three quantiles, $\tau \in \{0.15, 0.5, 0.85\}$, to illustrate performance for a low, central, and upper quantile.

The basic argument for 401(k) eligibility being a valid instrument for participation is that eligibility is determined by employers and so may plausibly be taken as an exogenous after conditioning on job relevant covariates. See, e.g., Abadie (2003) for further discussion of the underlying exclusion restriction. Of course, there are reasons that one might worry that the exclusion restriction does not hold perfectly. For example, one might conjecture that firms that offered 401(k) plans were attractive to employees who prefer savings for other, unobserved, reasons. Motivated by such concerns, Conley et al. (2012) explore sensitivity of linear IV estimates of the effect of 401(k) participation on financial assets. Our analysis extends this line of work by investigating the sensitivity of quantile treatment effect estimates. We also note that examining quantile effects may be of substantive economic interest given the strong asymmetry of financial asset holdings and potential interest in the effect of 401(k) plans on savings for those at different points of the wealth distribution.

[11] The covariates are income, a quadratic in age, family size, four indicators of education categories, marital status, two-earner status, defined benefit pension status, IRA participation, and home ownership. For more details about the data, see, e.g., Abadie (2003) and Chernozhukov and Hansen (2004).

Of course, implementing PGMM requires specification of priors for the structural parameters, θ_τ, and plausibility terms, μ_τ. In this example, we use the same diffuse prior for each θ_τ—$\theta_\tau \sim \mathcal{N}\left(0, \frac{10^{10}}{5} I_{14}\right)$—for all reported results. Given the magnitude of the outcome variable and units of the input variables,[12] this prior seems to be extremely uninformative.

The more delicate choice is the prior over the local misspecification parameter μ_τ. As in the previous example, we assess sensitivity to this choice by considering both zero-mean Gaussian priors and zero-mean uniform priors for each of our three values of τ. We set baseline priors using a stylized model for misspecification. Specifically, we construct a bound on the moment conditions, denoted δ_τ, under the assumption that any misspecification arises from a direct effect of D_t on Y_t, capped at 2000 dollars in absolute value. To simulate this bound, we compute

$$\max_{\gamma \in \{-2000, 2000\}} \left| \frac{1}{T} \sum_{t=1}^{T} (1, D_t, W_t^\top)^\top \left(\tau - \mathbf{1}\left(\epsilon_t + D_t \gamma \leqslant 0\right)\right) \right|,$$

where the ϵ_t are generated as the residuals from the linear IV analog of our quantile models. Using the resulting δ_τ, we define priors for μ_τ as $c \cdot \mathcal{N}\left(0, \mathrm{diag}(\delta_\tau/3)^2\right)$ or as uniform priors over $[-c\delta_\tau/3, c\delta_\tau/3]$. To explore varying degrees of prior concentration, we consider $c \in \{0, 0.5, 0.9, 1.0\}$, where $c = 0$ corresponds to the dogmatic prior that maintains correct specification.

Figure 2 displays marginal quasi-posteriors for both $\beta_{X,\tau}$ and the component of μ_τ associated with the IV moment condition, denoted by $\mu_{D,\tau}$, under the Gaussian prior for μ_τ with $c = 1$.[13] For comparison, we also provide the marginal quasi-posterior for $\beta_{X,\tau}$ under the assumption of correct specification ($c = 0$) in dashed curves. We see that the quasi-posteriors for the quantile effect obtained under correct specification concentrate over positive values for each value of τ, suggesting a robust positive effect of 401(k) participation on net financial assets. The quasi-posteriors under correct specification are suggestive of larger quantile treatment effects at higher quantiles.

Looking at the PGMM quasi-posteriors, we see that allowing for departures from correct specification according to our prior leads to substantially more diffuse quasi-posteriors. For $\tau = 0.15$ and $\tau = 0.85$, the quasi-posterior for $\beta_{X,\tau}$ now places substantial mass on both large positive and large negative values. This implies that, once we allow for plausible violations of the exclusion restriction—consistent with the instrument having up to a \$2,000 direct effect on savings—it becomes difficult to draw reliable conclusions about the lower and upper quantile treatment effect of 401(k) participation. The resulting intervals for low and high quantiles from "PGMM-g" in Figure 3 span from -7.88 to 18.09 and -13.66 to 33.23 (in units of 10^3), respectively. In contrast, the quasi-posterior for the median remains concentrated over positive values, suggesting a relatively robust positive median treatment effect of 401(k) participation.

Interestingly, Figure 2 reveals that the marginal quasi-posterior distributions of $\mu_{D,\tau}$ differ noticeably from the prior. In particular, the quasi-posterior for $\tau = 0.15$ ($\tau = 0.85$) is shifted

[12]For example, the 0.15 and 0.85 quantiles of Y_t are approximately -2751 and $36,303$, respectively.

[13]We provide quasi-posterior plots for the remaining settings in Figures 6-11 in the Appendix.

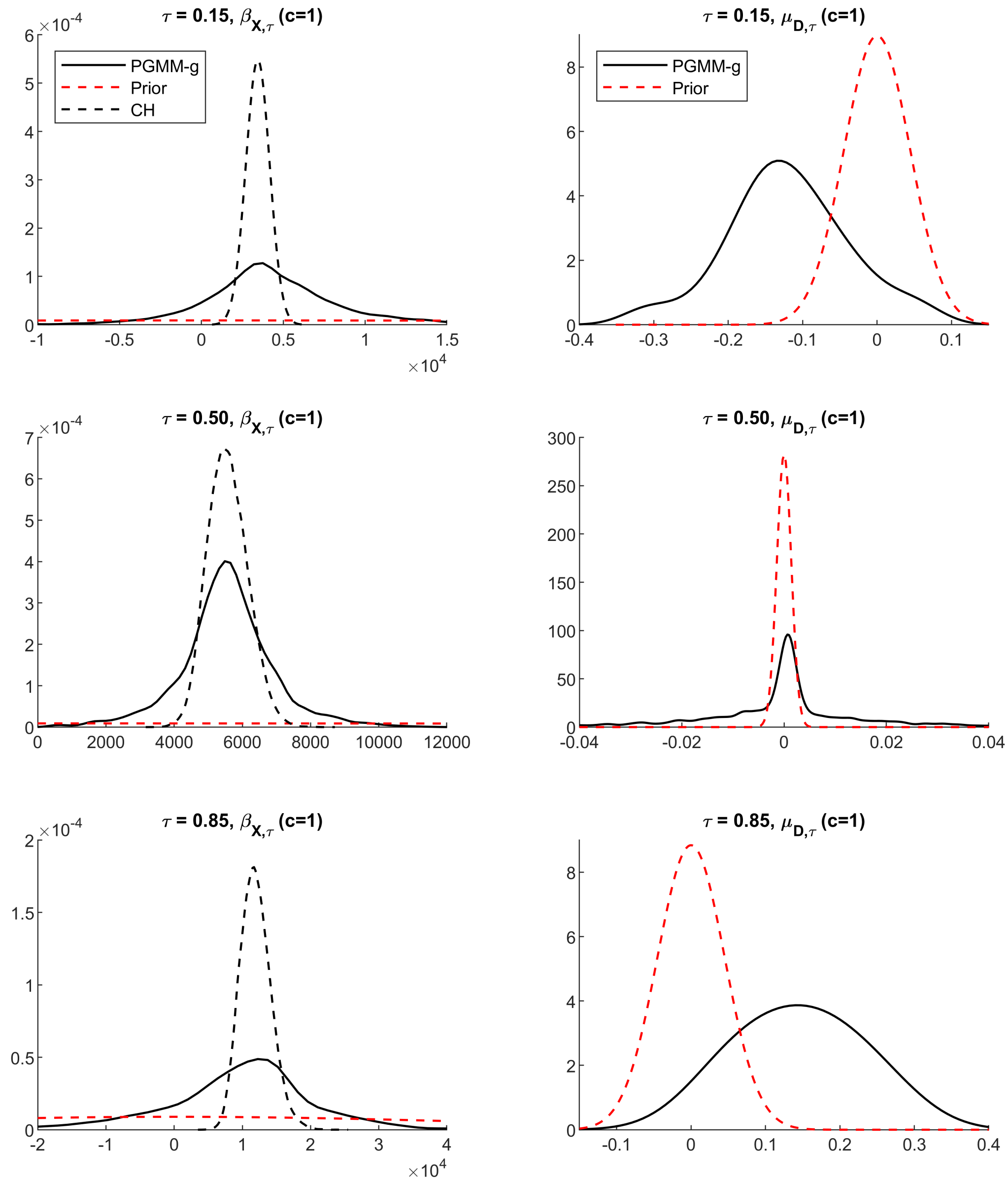

FIGURE 2. CH (dashed black curves, $\mu \equiv 0$) and PGMM marginal quasi-posteriors (solid black curves) for $\beta_{X,\tau}$ and $\mu_{D,\tau}$. Examples shown use Gaussian priors over μ that vary with τ with $c = 1$ as described in the main text. The red dashed curves represent the marginal prior density curves for the displayed parameters.

to the left (right) relative to the prior. The quasi-posterior for the median remains centered approximately over the prior center, but has substantially thicker tails than the marginal prior. This pattern suggests that deviations from the baseline model are more likely at the lower and upper quantiles, indicating that the moment conditions for the tails are "less plausible"—in the sense that their quasi-posteriors are not centered at zero—than those for the median. We find it interesting that, at least viewed through the lens of the marginal quasi-posteriors over $\mu_{D,\tau}$, the combination of data and moments leads to updating in the direction of model misspecification. That is, the marginal quasi-posterior over the sensitivity term for the IV moment restriction is less concentrated around zero (correct specification) than the initial prior. We take this updating as further evidence that a researcher should be hesitant to trust results that dogmatically maintain correct specification in this example.

Figure 3 reports 95% highest quasi-posterior density intervals constructed using the PGMM method under our full set of prior specifications for μ_τ. In addition to the PGMM intervals obtained from simulating the full quasi-posterior, we also report intervals based on the local limiting approximation described in Section 2.3 and Theorem 1, as well as frequentist intervals constructed using the method of Armstrong and Kolesár (2021) (AK). The AK intervals are derived under a local misspecification framework in which the true parameter value $\theta_{\tau,X}$ is assumed to satisfy $\sqrt{T}m(\theta_{\tau,X}) \in \mathscr{C}_\tau$. For comparability, we define the restriction set $\mathscr{C}_\tau$ to match the support of the corresponding uniform priors used in PGMM-u for a given constant c:

$$\mathscr{C}_\tau = \left\{\sqrt{T}\cdot \mathrm{diag}(c\delta_\tau/3)a : a \in \mathbb{R}^q,\ |a|_\infty \leq 1\right\},$$

where $|\cdot|_\infty$ denotes the ℓ_∞ norm. The parameter c in Figure 3 thus has a different interpretation for the different methods. For PGMM with a Gaussian prior on μ_τ (PGMM-g), it determines the standard deviations of the prior distribution. For PGMM with a uniform prior (PGMM-u), it sets the upper and lower bounds of the uniform prior support. For the local approximation intervals (Local Approx), it indexes the scale of the local Gaussian prior $\mathcal{N}(0, \Lambda_c/T)$, where $\Lambda_c = Tc^2 \cdot \mathrm{diag}((\delta_\tau/3)^2)$, consistent with the PGMM-g case. For the AK intervals, $c\delta_\tau$ parameterizes the local misspecification set $\mathscr{C}_\tau$ as defined above.

As shown in Figure 3, the results for the lower and upper quantiles appear relatively sensitive to both the value of c and the method used to obtain the interval estimate. As expected, the AK intervals—which are designed to ensure asymptotic frequentist coverage under worst-case local misspecification—are strictly wider than PGMM-u intervals in all cases. For these lower and upper quantiles, we see that the AK intervals are much wider than the corresponding PGMM intervals and, interestingly, are tracked relatively closely by the intervals obtained from the local approximation to the quasi-posterior. An interesting feature of this example is that the moment corresponding to plausibility term $\mu_{D,\tau}$ has natural support restrictions. The full quasi-Bayes procedure updates such that values that violate these support restrictions have very little posterior mass. This updating does not occur in either the local approximation or the AK intervals, which may explain some of the discrepancy between the procedures, especially for larger values of c.

In contrast, the intervals for the median effect are notably more stable across methods. All approaches yield similar interval estimates, and the lower bounds remain above zero even under relatively diffuse priors on the misspecification term. This stability suggests that inference

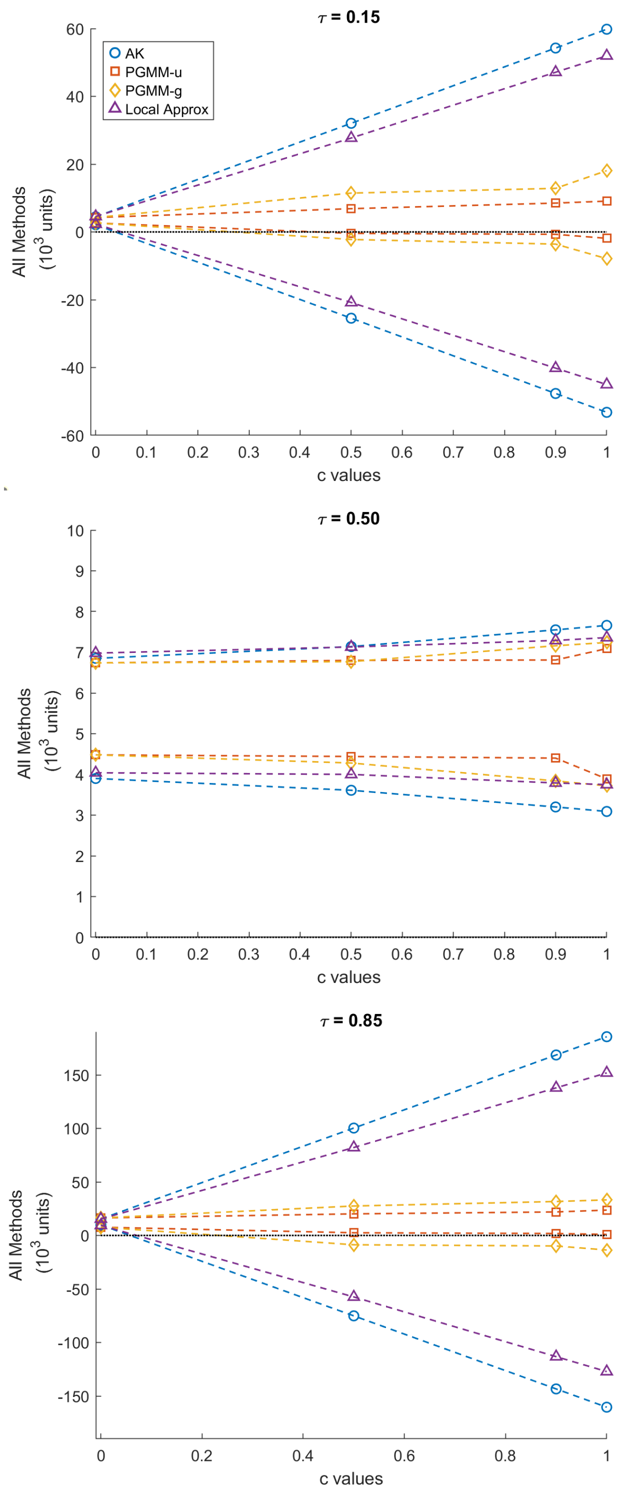

FIGURE 3. This figure shows the 95% intervals constructed for the treatment effect parameter, $\beta_{X,\tau}$, using the limiting approximation indicated by Theorem 1 (Local Approx), PGMM (PGMM-u denotes the cases with uniform priors on μ while PGMM-g denotes the cases with Gaussian priors on μ_τ) and AK (AK) for IVQR with $\tau = 0.15, 0.50, 0.85$ and $c = 0, 0.5, 0.9, 1.0$.

about the median treatment effect is more robust to the specification of priors and the choice of estimation method.

As in the previous example, we find that examining quasi-posteriors under non-dogmatic priors on the degree of misspecification offers valuable insight into the identification and plausibility of the estimated effects. Estimates of quantile effects in the upper and lower tails are relatively sensitive to assumptions about model specification, with this sensitivity manifesting as instability across methods and prior choices. In contrast, the estimated median effects appear considerably more robust, yielding qualitatively similar results across all considered specifications. Finally, the updating from the prior over μ_τ to the quasi-posterior is particularly informative. In all cases, the quasi-posteriors place more mass away from $\mu_\tau = 0$ than the prior, indicating quasi-posterior evidence against correct specification. This shift suggests that researchers should be cautious about imposing the assumption of correct specification too rigidly in this setting.

4. Theoretical results

This section is organized as follows. We first define notation in Subsection 4.1. Subsection 4.2 introduces a Bayesian optimal decision-theoretic motivation for the procedure. Subsection 4.3 presents a Bernstein–von Mises (BvM) theorem in a fixed-dimensional setting under local misspecification and establishes theoretical coverage guarantees for the highest quasi-posterior region. Subsection 4.4 extends the BvM result to the high-dimensional case. Finally, subsection 4.5 provides a frequentist justification for the coverage of the Bayesian credible set.

4.1. Notation.

For a vector $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ and $q > 0$, we denote $|v|_q = \left(\sum_{i=1}^d |v_i|^q\right)^{1/q}$, $|v|_\infty = \max_{1\le i\le d} |v_i|$, and $\|v\| = |v|_2$. For a vector v and a conformable non-negative definite matrix A, define $\|v\|_A := \sqrt{v^\top A v} \ge 0$. For two positive number sequences (a_T) and (b_T), we say $a_T \lesssim b_T$ (resp. $a_T \asymp b_T$) if there exists $C > 0$ such that $a_T/b_T \le C$ (resp. $1/C \le a_T/b_T \le C$) for all large T. We denote $a_T \ll b_T$ if $a_T/b_T \to 0$ as $T \to \infty$, and write $a_T \gg b_T$ if $a_T/b_T \to \infty$ as T diverges. Denote the total variation of moments (TVM) norm of κ for a real-valued measurable function g on Θ by $\|g\|_{TVM(\kappa)} = \int_{h\in\Theta}(1 + \|h\|^\kappa)|g(h)|dh$ for $\kappa > 0$. We use $\propto$ to denote "proportional to". We use the subscript p to denote statements with respect to the outer measure $\mathbb{P}^*$ of a given probability $\mathbb{P}$. We use $\to_d$ to denote convergence in distribution. We set (X_T) and (Y_T) as two sequences of random variables. Write $X_T = O_p(Y_T)$ if $\forall \epsilon > 0$, there exists $C > 0$ such that $\mathbb{P}^*(|X_T/Y_T| \le C) > 1 - \epsilon$ for all large T. We denote $X_T = o_p(Y_T)$ if $X_T/Y_T \to_p 0$ as $T \to \infty$. We limit ourselves to situations in which, given μ, observations are a random sample from a distribution $\mathbb{P}_\mu$ with $\mathbb{P}_\mu$ being the conditional law of the random sample given μ; and probability statements under $\mathbb{P}$ are made relative to the joint distribution of the random sample and μ, given a fixed latent distribution F_μ over μ.

4.2. **Link to Bayes optimal decisions.** Given the quasi-posterior, we can formulate optimal decisions under the quasi-posterior by minimizing the quasi-posterior expected risk. Specifically, let $\ell(\theta,\mu,d)$ be a loss functions that depend on the parameters θ,μ and a decision $d\in\mathscr{D}$. Policymakers may want to choose a decision d to minimize the loss $\ell\left(\theta,\mu,d\right)$ that depends on both θ and μ. The expected loss-minimizing decision under the quasi-posterior, denoted by $s_T(p_T)$, then takes the usual form:

$$s_T(p_T)\in\arg\min_{d\in\mathscr{D}}\int \ell(\theta,\mu,d)\,p_T(\theta,\mu)\,d\mu d\theta. \tag{6}$$

We note that this framework encompasses the likely leading setting where loss depends only upon θ, where loss is $\ell(\theta,d)$, as a special case.

These quasi-Bayes decision rules can be motivated as approximating fully Bayesian rules. Andrews and Mikusheva (2022) show that in the weak identification cases, the quasi-posterior based on the continuously updated GMM can be obtained as the limit of a sequence of posteriors under proper priors, and the resulting quasi-Bayes decision rule can correspond to the pointwise limit of the sequence of Bayes decision rules.

In the case where parameters are low-dimensional, the results of Andrews and Mikusheva (2022) can readily be adapted to the PGMM framework. Specifically, we have that, under regularity conditions, e.g., Assumptions 1 and 3.ii), the process $\sqrt{T}\widehat{m}(\cdot)-\sqrt{T}m(\cdot)$ converges in distribution to a mean-zero Gaussian process with covariance function $\Sigma(\cdot,\cdot)$ and mean function satisfying $m(\theta(\mu))=\mu$ on $\mu\in\Gamma$. It then follows that we can construct a likelihood as in Andrews and Mikusheva (2022) by properly substituting their parameter θ^* with the pair $(\theta(\mu),\mu)$. As a result, the optimal quasi-Bayesian decision rule under model misspecification retains the desirable properties established by their analysis. We provide the supporting technical details in Supplementary Appendix SA-2.

4.3. **Gaussian quasi-posterior approximation under local misspecification.** This subsection considers a local misspecification setting in which the prior on μ is Gaussian with variance Λ/T. We show that, under this specification, the quasi-posterior distribution is asymptotically Gaussian and coincides with the results in Chernozhukov and Hong (2003) in the special case where $\Lambda=0$—i.e., in the case that $\mu\equiv 0$. The formal result in this section, Theorem 1, provides additional justification for the arguments presented in Section 2.

We start by presenting the technical assumptions under which we establish the Gaussian approximation. Throughout this section, we set $\mu_0=0$ and define

$$G(\theta(\mu))=\frac{\partial\mathbb{E}_{P_\mu}\left[g(Z_t,\theta(\mu))\right]}{\partial\theta(\mu)}.$$

We also let $\theta(\mu_0)=\theta_0$, $A_{\theta(\mu_0)}=A_{\theta_0}$, and

$$\Pi_T(\theta,\mu_0)\propto\exp\left(-\frac{T}{2}\|\theta-\widehat{\theta}\|^2_{G(\theta_0)^\top A_{\theta_0}G(\theta_0)}\right).$$

Assumption 1 (Plausibility characteristic). *The plausiblility characteristic* $\mu \in \mathcal{M} \subset \mathbb{R}^q$. $\Gamma \subseteq \mathcal{M}$ *is a set containing values for* μ *such that* $\theta(\mu)$ *satisfying* $m(\theta(\mu)) = \mu$ *exists and is unique. Each* $\theta(\mu)$ *belongs to the interior of a compact convex subset* Θ *of the Euclidean space* $\mathbb{R}^k$.

Assumption 1 defines the support of the plausibility characteristic and, importantly, a set Γ within the support such that moment condition has a unique solution for each value of $\mu \in \Gamma$. In Assumption 5 below, we require that the prior for μ has positive mass over at least one point in Γ which ensures the quasi-Bayes procedure is (asymptotically) relatively well-behaved.

Assumption 2 (GMM estimator). *Assume* $m(\theta)$ *is first order differentiable in* $\theta \in \Theta$. *Let*

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \widehat{m}(\theta)^\top \widehat{A}_{T,\theta} \widehat{m}(\theta)$$

be the GMM estimator using weighting matrix $\widehat{A}_{T,\theta}$. *Assume* $\widehat{\theta}$ *has expansion*

$$\widehat{\theta} = \theta(\mu_0) + J_A(\theta(\mu_0))^{-1} \Delta_T\left(\theta(\mu_0)\right) + o_p(1/\sqrt{T}),$$

where $J_A(\theta(\mu_0)) = G(\theta_0)^\top A_{\theta_0} G(\theta_0)$ *and* $\Delta_T(\theta_0) = -G(\theta_0)^\top A_{\theta_0}(\widehat{m}(\theta_0) - \mu_0)$.

Assumption 3 (Expansion). *i) Assume* $J_A(\theta)$ *is positive definite for all* $\theta \in \Theta$, *and* $J_A(\theta)$ *is continuous in* θ. *Further, assume* $G(\theta)$ *and* $\Omega(\theta)$ *are continuous and full rank for all* $\theta \in \Theta$, $\mu \in \mathcal{M}$. *ii) Assume*

$$\Delta_T\left(\theta(\mu)\right)/\sqrt{T} = -\sqrt{T} G(\theta(\mu))^\top A_{\theta(\mu)}(\widehat{m}\left(\theta(\mu)\right) - \mu) \to_{\mathrm{d}} \mathcal{N}\left(0, \tilde{V}\left(\theta(\mu)\right)\right),$$

where

$$\tilde{V}(\theta(\mu)) = G(\theta(\mu))^\top A_{\theta(\mu)} \Omega(\theta(\mu)) A_{\theta(\mu)} G(\theta(\mu)).$$

iii)

$$\Omega(\theta(\mu)) = \lim_{T\to\infty} Var\left(\sqrt{T}\left(\widehat{m}(\theta(\mu)) - m(\theta(\mu))\right)\right).$$

Assumption 4 (Modulus of continuity and identification.). *Let*

$$r_T(m,\theta) = \sqrt{T}\, \|(\widehat{m}(\theta) - \widehat{m}(\theta_0)) - (\mathbb{E}\widehat{m}(\theta) - \mathbb{E}\widehat{m}(\theta_0))\|.$$

For a sufficiently small positive constant $\delta > 0$, *assume*

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

and $r(\delta) \to_p 0$ *if* $\delta \to 0$. *Further, assume*

$$\inf_{\theta:\|\theta-\theta_0\|\geq\delta} \|(\widehat{m}(\theta) - \widehat{m}(\theta_0))\| \geq \delta.$$

Assumption 2 refers to the properties of the GMM estimator with a weighting matrix that incorporates prior uncertainty as outlined in Section 2.3. It specifically imposes that the resulting GMM estimator has a linear expansion dominated by its leading term. Assumption 3 then addresses the asymptotic behavior of the leading term in the GMM estimator. Assumptions 2 and 3 are analogous to standard assumptions that align, for example, with Assumption 4 and conditions (ii) and (iii) in Proposition 1 in Chernozhukov and Hong (2003). Assumption 4 is a type of modulus of continuity assumption similar to condition (iv) in Proposition 1 in

Chernozhukov and Hong (2003), which is employed to handle non-smooth criterion functions. It requires a remainder term to be bounded within a neighborhood of θ_0 and holds when moments are sufficiently smooth. The last condition in Assumption 4 ensures that θ_0 is asymptotically well-identified.

The next assumption imposes restrictions on the prior that are sufficient for verifying approximate normality of the quasi-posterior.

Assumption 5 (Prior)**.** *$\pi(\mu,\theta) = \pi(\mu)\pi(\theta)$, where $\pi(\mu)$ is a Gaussian prior centered at $\mu_0 \in \Gamma$ with covariance matrix Λ/T, and $\pi(\theta)$ is bounded and continuously differentiable around an open ball of $\theta_0 \in \Theta$. $\lambda_{\max}(\Lambda) \ll T$, and $1 \lesssim \lambda_{\max}(\Lambda)/\lambda_{\min}(\Lambda) \lesssim 1$, where $\lambda_{\min}(\Lambda)$ and $\lambda_{\max}(\Lambda)$ denote the minimum and the maximum eigenvalue of a matrix Λ respectively.*

The main substantive restriction of Assumption 5 is that the prior for μ is Gaussian with variance of the same order as sampling variation. We further impose that the prior variance is full rank and that, *a priori*, θ and μ are independent. The requirement that Λ be of full rank can be relaxed.[14]

We now present the first main theorem, which shows that the quasi-posterior density $p_T(\theta)$ converges in the TVM norm to a Gaussian density. With slight abuse of notation, we denote $\bar{p}_T(\widehat{m}(\theta)) = p_T(\theta)$ and obtain $\bar{p}_T\left(\widehat{m}(\widehat{\theta}) - G(\theta_0)(\widehat{\theta}-\theta)\right)$ by replacing $\widehat{m}(\theta)$ in $\bar{p}_T(\widehat{m}(\theta))$ with $\widehat{m}(\widehat{\theta}) - G(\theta_0)(\widehat{\theta}-\theta)$.

Theorem 1. (Convergence in TVM norm). Under Assumptions 1 − 5, for any $0 \le \kappa < \infty$,

$$\begin{aligned}
&\left\| \bar{p}_T\left(\widehat{m}(\widehat{\theta}) - G(\theta_0)(\widehat{\theta}-\theta)\right) - \Pi_T(\theta,\mu_0) \right\|_{TVM(\kappa)} \\
&\qquad = \int_{\theta\in\Theta} \left(1 + \|\theta-\theta_0\|^{\kappa}\right) \left| \Pi_T(\theta,\mu_0) - \bar{p}_T\left(\widehat{m}(\widehat{\theta}) - G(\theta_0)(\widehat{\theta}-\theta)\right) \right| \mathrm{d}\theta \to_{\mathrm{p}} 0, \text{ and} \\
&\left\| p_T(\theta) - \Pi_T(\theta,\mu_0) \right\|_{TVM(\kappa)} = \int_{\theta\in\Theta} \left(1 + \|\theta-\theta_0\|^{\kappa}\right) \left| p_T(\theta) - \Pi_T(\theta,\mu_0) \right| \mathrm{d}\theta \to_{\mathrm{p}} 0.
\end{aligned}$$

Proof. See Appendix 7.1. □

Theorem 1 demonstrates that $p_T(\theta)$ can be asymptotically approximated by a Gaussian density function $\Pi_T(\theta,\mu_0)$ under a sequence of Gaussian priors over misspecification that concentrate at the same rate as sampling error. Further, the theorem confirms the expected result that $p_T(\theta)$ concentrates in a $1/\sqrt{T}$ neighborhood of θ_0 under local misspecification. This result differs from the related approximation result in Chernozhukov and Hong (2003) in that the quasi-posterior depends on the prior for the plausibility characteristic, even asymptotically. We do note that the approximation result reproduces the result from Chernozhukov and Hong (2003) under $\mu_0 = 0$

[14] For example, let $B \in \mathbb{R}^{q\times\tilde{q}}, \tilde{q} < q,$, and let $\Lambda_x \in \mathbb{R}^{\tilde{q}\times\tilde{q}}$ be full rank. Consider the prior for μ generated from $\mu = Bx, x \sim \mathcal{N}\left(0,\ T^{-1}\Lambda_x\right)$. Following the same arguments as used to establish Theorem 1, we can establish the quasi-posterior density $p_T(\theta)$ is, for large T, approximately Gaussian with covariance matrix $A_\theta = \Omega(\theta)^{-1} - \Omega(\theta)^{-1} B\left(\Lambda_x^{-1} + B^\top\Omega(\theta)^{-1}B\right)^{-1} B^\top\Omega(\theta)^{-1}$.

and $\Lambda \to 0$—i.e., when the prior over misspecification concentrates more quickly than sampling error.

We note that Theorem 1, along with Theorems 2-4 presented below, could be established without requiring the data stream $\{Z_t\}_{t=1}^{T}$ to be i.i.d. Rather, we could work with moments defined as $m_T(\theta(\mu)) = T^{-1}\sum_{t=1}^{T}\mathbb{E}_{\mathbb{P}_{\mu,t}}[g(Z_t,\theta(\mu))] = \mu$ where $\mathbb{P}_{\mu,t}$ is the marginal distribution for Z_t. Within this structure, results could be established under suitable restrictions on dependence and heterogeneity. We do not pursue this direction formally to avoid further complicating notation.

4.4. **General quasi-posterior approximation results.** This section discusses an extension of Theorem 1 by allowing relatively general choice of prior for μ. The key result is that, as in the previous section, the prior for μ matters even in the limit. However, under a general prior structure, we do not obtain a limiting Gaussian approximation. Rather, we have that, *conditional* on μ, the limiting approximation is Gaussian. Thus, the limiting approximation to the posterior is a Gaussian mixture where mixture weights depend heavily on the prior. We establish the formal results under sequences that allow the dimensions k and q to grow with the sample size T, which offers a technical extension of some results even in the case where a dogmatic prior is placed over μ.

To accommodate a broader family of weighting matrices, we allow the GMM-type criterion that serves to define our quasi-posterior, $Q_T(\theta,\mu)$, to be formed with any positive-definite weight matrix $\widehat{W}_T(\theta)$. That is, we now consider

$$Q_T(\theta,\mu) = -T\left(\widehat{m}(\theta)-\mu\right)^\top \widehat{W}_T(\theta)\left(\widehat{m}(\theta)-\mu\right),$$

where setting $\widehat{W}_T(\theta) = \widehat{\Omega}_T(\theta)^{-1}$ corresponds to the leading case discussed in previous sections. Within this more general formulation, we use $W(\theta)$ to denote the population analog of $\widehat{W}_T(\theta)$ in the same manner that $\Omega(\theta)$ serves as the population counterpart of $\widehat{\Omega}_T(\theta)$.

In stating our formal results in this section, we make use of additional notation. Let $h(\theta,\mu) = G(\theta(\mu))(\theta-\theta(\mu))$, for $\mu \in \Gamma$, and $h(\theta,\mu) = G(\theta(\mu))(\theta-\theta(\nu(\mu))))$, for $\mu \not\in \Gamma$, where we let $\nu(\mu)\in\Gamma$ be the closest point of μ for $\mu \in \mathscr{M}$ according to $\|.\|$. Let the $\varepsilon/\sqrt{T}$ expansion of $\{(\theta(\mu),\mu):\mu\in\Gamma\}$ be defined within $B_\varepsilon = \left\{(\theta,\mu):\sqrt{T}\|h(\theta,\mu)\| \le \varepsilon, \sqrt{T}\|\mu-\nu(\mu)\| \le \log(T)\varepsilon, \theta\in\Theta, \mu\in\mathscr{M}\right\}$ with $\varepsilon \asymp \sqrt{k}\log T$. B_ε is thus a $\varepsilon/\sqrt{T}$-set containing elements (θ,μ) closely around the plausible pairs $(\theta(\mu),\mu)$ with $\mu\in\Gamma$. Let

$$V_T(h(.),\theta,\mu) = 2Th^\top W(\theta(\mu))(\widehat{m}(\theta(\mu))-\mu) + Th^\top W(\theta(\mu))h + C(\mu),$$

where $C(\mu) = -2\log(\pi(\theta(\mu))) + T(\widehat{m}(\theta(\mu))-\mu))^\top W(\theta(\mu))(\widehat{m}(\theta(\mu)-\mu)$ and we abbreviate $h(\theta,\mu)$ as h. For $\mu\in\mathscr{M}/\Gamma$, we define $V_T(h(.),\theta,\mu)$ by using $\theta(\nu(\mu))$ in place of $\theta(\mu)$.

We now state sufficient conditions for establishing our limiting approximation to the quasi-posterior.

Assumption 6 (Identification, smoothness, and tails). *Assume the following: (i) $m(\theta)$ is first order differentiable. (ii) For any $\theta \in \Theta$, $G(\theta)$ and $W(\theta)$ have singular values bounded from below and above. (iii) There exists a positive constant $C > 0$, $\sup_{1\le j\le q,(\theta,\mu)\in B_\varepsilon} \mathbb{E}\left[\|e_j^\top(g(Z_t,\theta)-\mu)\|^2\right] < C$. (iv) For $\theta,\theta' \in \Theta$, $\mu \in \Gamma$, $\|G(\theta)-G(\theta')\| \lesssim \|\theta-\theta'\|$, $\|W(\theta)-W(\theta')\| \lesssim \|\theta-\theta'\|$, and $\|\widehat{W}_T(\theta)-W(\theta)\| = o_p(1)$.*

Assumption 6 imposes regularity conditions that are sufficient for good behavior of the limiting criterion function. Importantly, these conditions ensure that, at fixed values of μ, θ is strongly identified through the restrictions on $G(\theta)$; see, e.g., Hansen et al. (2010).

The next assumption restricts priors, importantly requiring that θ and μ are *a priori* independent and that the marginal priors place positive mass uniformly over the corresponding parameter space.

Assumption 7 (Prior). *Let $\mathcal{M}$ be an interior of a compact convex subset of $\mathbb{R}^q$. Assume the prior density $\pi(\theta,\mu) = \pi(\theta)\pi(\mu)$ with $\pi(\mu) > 0$ for $\mu \in \mathcal{M}$ and $\pi(\theta) > 0$ for $\theta \in \Theta$. Assume $\pi(\theta)$ is bounded and continuously differentiable on Θ.*

Before stating the next assumption, define

$$R_T(\theta,\mu) = \frac{1}{2}(Q_T(\theta,\mu) + V_T(h(.),\theta,\mu)) + \log(\pi(\theta)).$$

Assumption 8 (Empirical process). *There exists $\varepsilon \asymp \sqrt{k}\log T$ such that the following properties hold with probability tending to 1:*

$$\frac{\sup_{\theta,\mu\in B_\varepsilon} T|R_T(\theta,\mu)|}{(\|\sqrt{T}h(\theta,\mu)\|^2 + k(\log T)^2)} = O_p\left(\frac{\sqrt{k}(\log T)^2}{\sqrt{T}} + \frac{q}{\sqrt{kT}}\right)$$

with $\frac{(\log T)((\log T)^2 k)^{(\kappa\vee 2)+1}}{T} \to 0$ and $\frac{k(\log T)^4 q^2}{T} \to 0$, and there exists a positive constant $\frac{1}{2} < C_0 \le 1$ such that for all θ,μ satisfying $\sqrt{T}h(\theta,\mu) \ge \epsilon$,

$$TR_T(\theta,\mu) \le -C_0\sqrt{T}\|\theta-\theta(\nu(\mu))\|_{J_W(\theta(\mu))}\varepsilon + \frac{C_0\varepsilon^2}{2} + \frac{T\|\theta-\theta(\nu(\mu))\|^2_{J_W(\theta(\nu(\mu)))}}{2} \tag{7}$$

with $J_W(\theta(\mu)) = G^\top(\theta(\mu))W(\theta(\mu))G(\theta(\mu)$.

The condition

$$\sup_{\theta,\mu\in B_\varepsilon} \frac{T|R_T(\theta,\mu)|}{\|\sqrt{T}h(\theta,\mu)\|^2 + k(\log T)^2} \to_p 0$$

arises from needing to control a modulus of continuity. It ensures the oscillatory behavior of the empirical process $R_T(\theta,\mu)$ is mild. While the condition is high level, Lemma 4 shows that this condition is satisfied with differentiable moments. Assumption 7 effectively imposes an identification requirement for large values of θ and a smoothness condition for smaller values of θ on the term $\sup_{(\theta,\mu)\in B_\varepsilon^c} TR_T(\theta,\mu)$. The assumption resembles the finite-sample bound in

Lemma A.16 of Spokoiny and Panov (2019), and it enables the derivation of a tail bound outside the ball B_ε using a Gaussian integral argument.

Now, define $N_T(\theta,\mu) = \frac{\exp\{-\frac{1}{2}[V_T(h(.),\theta,\mu)]\}\pi(\mu)}{\int_{\mu\in\mathcal{M}}\int_{\theta\in\Theta}\exp\{-\frac{1}{2}[V_T(h(.),\theta,\mu)]\}\pi(\mu)d\theta d\mu}$. Under the stated assumptions, we obtain the following Bernstein-von Mises-type result establishing that the quasi-posterior is asympotically approximated by $N_T(\theta,\mu)$.

Theorem 2. Under Assumptions 1, 6-8, we have,

$$\left\| p_T(\theta,\mu) - N_T(\theta,\mu) \right\|_{TVM(\kappa)} \equiv \int_{\mu\in\mathcal{M}}\int_{\theta\in\Theta} \left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\left|p_T(\theta,\mu)-N_T(\theta,\mu)\right| \mathrm{d}\theta d\mu \to_{\mathrm{p}} 0. \tag{8}$$

Proof. See Appendix 7.2 □

The above theorem indicates that, conditional on fixed values of μ, the quasi-posterior distribution of θ can be well approximated by a Gaussian distribution, thus facilitating practical inference via conditional sampling, as demonstrated in Theorems 3-4. In contrast to Theorem 1, Theorem 2 relaxes the prior specification on μ by not imposing a Gaussian prior, thereby extending the applicability of the result. While the joint limiting distribution $N_T(\theta,\mu)$ is not Gaussian in general, it becomes Gaussian when conditioning on μ. This insight has practical implications: one may select a representative set of μ values, compute the corresponding conditional quasi-posterior distributions of θ, and aggregate the highest quasi-posterior density regions. This approach mirrors the strategy employed by Conley et al. (2012) for constructing robust inference under partial identification.

While quasi-Bayes posteriors can be motivated as providing approximate Bayesian uncertainty quantification, we also note that we can use posterior intervals to provide directly to provide approximate frequentist coverage within a two-stage sampling regime.

Specifically, let $PR_T(\alpha)$ denote the $(1-\alpha)\%$ $(0<\alpha<1)$ highest quasi-posterior density region for θ obtained from the quasi-posterior $p_T(\theta)$.[15] Lemma 1 shows that $PR_T(\alpha)$ asymptotically provides valid weighted average frequentist coverage in large samples if we envision a world where nature draws μ from the prior $\pi(\mu)$, in which case $\pi(\mu)$ coincides with the fixed latent distribution F_μ in the data generating mechanism.

Lemma 1. (Weighted average coverage rate of $PR_T(\alpha)$) Assume that $\pi(\mu)$ coincides with F_μ and that Assumptions 1, 6-8 hold. Let $\partial F_\mu(u)/\partial\mu = f_\mu(u)$ and fix $\alpha\in(0,1)$. Assume that $W(\theta(\mu))^{-1} = \Omega(\theta(\mu))$ and the distribution of $\widehat{m}(\theta(\mu))-\mu$ under $\mathbb{P}_\mu$ can be well approximated by a Gaussian distribution with mean zero and covariance matrix $\Omega(\theta(\mu))$ as $T\to\infty$. Then $PR_T(\alpha)$ satisfies the following π-weighted average coverage rate, which also corresponds to the coverage rate under

[15] For a positive constant c and density $p_T(\theta)$, $PR_T(\alpha)=\{\theta\in\Theta: p_T(\theta)>c\}$ such that $\int_{PR_T(\alpha)} p_T(\theta)d\theta = 1-\alpha$.

$\mathbb{P}$, in large samples:

$$\int_{\mu} \mathbb{P}_{\mu}(\theta(\mu) \in PR_T(\alpha))\pi(\mu)d\mu = \mathbb{P}(\theta(\mu) \in PR_T(\alpha)) \approx 1-\alpha.$$

Proof. See Appendix 7.3. □

4.5. **Using quasi-posteriors to provide frequentist inference under support restrictions.** In this section, we provide an approach to obtain regions for θ that deliver valid frequentist coverage guarantees under support restrictions over μ. The regions are constructed by taking unions of quasi-posterior credible regions for fixed values of μ. Frequentist validity of this approach relies on properties of quasi-posterior intervals obtained from the posterior distribution $p_T(\theta,\mu)$ established in Theorems 3-4 in this section.

In the following, we suppose that one is interested in a continuously differentiable function $\eta(\theta): \mathbb{R}^k \to \mathbb{R}$. To state our results, we define the following quantities at fixed, given values of μ:

$$J_{\Omega,W}\left(\theta(\mu)\right) = G(\theta(\mu))^{\top} W(\theta(\mu))\Omega(\theta(\mu))W(\theta(\mu))G(\theta(\mu)),$$

$$U_T(\mu) = J_W^{-1/2}\left(\theta(\mu)\right)\Delta_{T,W}\left(\theta(\mu)\right),$$

$$\Delta_{T,W}\left(\theta(\mu)\right) = G(\theta(\mu))^{\top} W(\theta(\mu))(\widehat{m}(\theta(\mu))-\mu).$$

We first introduce a high level assumption for an estimator of $\theta(\mu)$ which is defined at a fixed value of μ. While we focus on quasi-Bayes estimators in this paper, we note that the estimator in this section can be relatively generic. For example, it could be a LTE estimator conditional on a value of μ— $\widehat{\theta}(\mu) = \operatorname{argmin}_{d\in\mathscr{D}}\int_{\theta}\ell(\theta,d)p_T(\theta,\mu)/p_T(\mu)d\theta$— or a CUE extreme estimator—$\widehat{\theta}(\mu) = \operatorname{argmin}_{\theta\in\Theta}Q_T(\theta,\mu)$—among many others

Assumption 9. *(Asymptotic normality) For a fixed $\mu\in\Gamma$, $\widehat{\theta}(\mu)$ admits the following linearization:*

$$\|(\widehat{\theta}(\mu)-\theta(\mu)) - U_T(\mu)\| = o_p\left(\frac{k}{\sqrt{T}}\right),$$

and

$$\sqrt{T}\sigma_{\eta,\mu}^{-1}(\partial\eta(\theta(\mu))/\partial\theta)^{\top}(\widehat{\theta}(\mu)-\theta(\mu)) \to_d \mathscr{N}(0,1),$$

where $\sigma^2_{\eta,\mu} = (\partial\eta(\theta(\mu))/\partial\theta)^{\top} J_W(\theta(\mu))^{-1} J_{\Omega,W}\left(\theta(\mu)\right) J_W(\theta(\mu))^{-1}(\partial\eta(\theta(\mu))/\partial\theta).$

The expansion assumed in Assumption 9 is readily justified for GMM extremum estimators; see, e.g., Corollary SA-1. The proof of Theorem 2 implies that, for any $\mu\in\Gamma$, the CUE extremum estimator admits the same first-order linearization as the LTE with symmetric loss functions analyzed in Chernozhukov and Hong (2003), when considering the quasi-posterior conditional on μ; see also Theorem 2 of Chernozhukov and Hong (2003) for the fixed-k case. When the weight matrix satisfies the generalized information equality (Equation 9), Assumption 9 further implies that the asymptotic variance of the leading term $(\partial\eta(\theta(\mu))/\partial\theta)^\top U_T(\mu)$ in the expansion of $\eta(\widehat{\theta}(\mu))-\eta(\theta(\mu))$ is given by

$$T^{-1}\left(\frac{\partial\eta(\theta(\mu))}{\partial\theta}\right)^\top J_\Omega\left(\theta(\mu)\right)^{-1}\left(\frac{\partial\eta(\theta(\mu))}{\partial\theta}\right).$$

We now state two theorems that make use of different features of quasi-posteriors obtained conditional on fixed values of μ to produce interval estimates for $\eta(\theta)$. The first main result in each theorem verifies that the resulting interval estimates have asymptotically correct frequentist coverage under the assumption that the fixed value of μ corresponds to the value of μ defining the conditional distribution from which data were realized. As a consequence, we can obtain frequentist confidence regions with correct coverage under the prior support condition that μ belongs to a known set $\mathscr{M}$ without requiring a completely specified prior by taking a union of confidence intervals produced at each $\mu\in\mathscr{M}$. This approach is analogous to the union of confidence intervals approach in Conley et al. (2012) and the approach outlined in Remark 3.3 of Armstrong and Kolesár (2021).

Theorem 3. (Frequentist properties of posterior quantiles) Assume that $\eta(\theta)$ has bounded derivatives such that for a positive constant c, $\|(\partial\eta(\theta(\mu))/\partial\theta)\|\le c$, and

$$\lim_{T\to\infty} J_\Omega\left(\theta(\mu)\right)J_{\Omega,W}\left(\theta(\mu)\right)^{-1}=I_{k\times k},\tag{9}$$

where $I_{k\times k}$ denotes the $k\times k$ identity matrix. Let $c_{\eta,T}(\alpha,\mu)$ be the α conditional quantile corresponding to the posterior density $p_T(\theta,\mu)$ given μ such that $\{\inf_{x\in\mathbb{R}}:F_{\eta,T}(x,\mu)\ge\alpha\}$, where the conditional quasi-posterior distribution function, $F_{\eta,T}(x,\mu)$, for each $\mu\in\Gamma$, is defined as

$$F_{\eta,T}(x,\mu)=\int_{\theta\in\Theta:\eta(\theta)\le x}p_T(\theta,\mu)/p_T(\mu)d\theta.$$

Under Assumptions 1, 6-9, for any $\alpha\in(0,1)$,

$$c_{\eta,T}(\alpha,\mu)-\eta(\widehat{\theta}(\mu))-q_\alpha\frac{\sqrt{\left(\partial\eta\left(\theta(\mu)\right)/\partial\theta\right)^\top J_\Omega\left(\theta(\mu)\right)^{-1}\left(\partial\eta\left(\theta(\mu)\right)/\partial\theta\right)}}{\sqrt{T}}=o_p\left(\frac{k}{\sqrt{T}}\right).$$

Let $\mathrm{CI}(\mu)=[c_{\eta,T}(\alpha/2,\mu),c_{\eta,T}(1-\alpha/2,\mu)]$, then

$$\lim_{T\to\infty}\mathbb{P}^*_\mu\left\{\eta(\theta(\mu))\in\mathrm{CI}(\mu)\right\}=1-\alpha,\tag{10}$$

and

$$\lim_{T\to\infty}\mathbb{P}^*\left\{\eta(\theta(\mu))\in\cup_{\mu'\in\mathscr{M}}\mathrm{CI}(\mu'),\forall\mu\in\Gamma\right\}\ge1-\alpha.\tag{11}$$

Proof. See Appendix 7.4. □

Theorem 3 verifies that quantiles obtained from the limiting conditional posterior density $N_T(\theta,\mu)/N_T(\mu)$ provide a valid frequentist approximation to the distribution of $\sqrt{T}(\eta(\widehat{\theta}(\mu))-\eta(\theta(\mu)))$. This coverage result relies on the generalized information equality, (9), and is consistent with findings for point-identified scalar parameters and partially identified models, as documented in Chernozhukov and Hong (2003) and Chen et al. (2018).

The first main result of Theorem 3, (10), verifies that posterior quantiles obtained from the quasi-posterior constructed conditional on fixed value of μ have asymptotically correct coverage under the corresponding conditional measure. (11), showing valid coverage of the union of intervals obtained under a support restriction, then immediately follows under the assumption that the value of μ under which the data were generated belongs to the specified support $\mathcal{M}$.

The results of Theorem 3 critically depend on choosing $\widehat{W}_T(\theta(\mu))$ such that (9) holds. Motivated by Theorem 4 of Chernozhukov and Hong (2003), the next result proposes an alternative procedure for using features of the quasi-posterior to construct intervals with frequentist coverage guarantees in settings where $\widehat{W}_T(\theta(\mu))$ is specified in such a way that (9) does not hold.

Theorem 4. (Frequentist properties of intervals based on Gaussian approximations) Suppose Assumptions 1, 6-9 hold and that $\eta(\theta)$ has bounded derivatives. Let c be a small enough positive constant such that $\int 1_{\{p_T(\mu)\leq c\}} f_\mu(\mu) d\mu = o_p(1)$. Let

$$\widehat{J}_T^{-1}\left(\widehat{\theta}(\mu)\right) = \int_\Theta T(\theta - \widehat{\theta}(\mu))(\theta - \widehat{\theta}(\mu))^\top \left[\frac{p_T(\theta,\mu)}{p_T(\mu)}\right] d\theta 1(\mu \in \{\mu \in \mathcal{M} : p_T(\mu) > c\}),$$

and assume that there exists an estimator $\widetilde{J}_{\Omega,W}\left(\theta(\mu)\right)$ such that

$$\|\widetilde{J}_{\Omega,W}\left(\theta(\mu)\right) J_{\Omega,W}\left(\theta(\mu)\right)^{-1} - I_{q\times q}\| \to_p 0.$$

We then have

$$\|\widehat{J}_T\left(\widehat{\theta}(\mu)\right) J_W^{-1}\left(\theta(\mu)\right) - I_{q\times q}\| \to_p 0$$

with $I_{q\times q}$ being the $q\times q$ identity matrix. Let

$$\tilde{c}_{\eta,T}(\alpha,\mu) \stackrel{\text{def}}{=} \eta(\widehat{\theta}(\mu)) + q_\alpha \cdot \frac{\sqrt{\left(\partial\eta\left(\widehat{\theta}(\mu)\right)/\partial\theta\right)^\top \widehat{J}_T\left(\widehat{\theta}(\mu)\right)^{-1} \widetilde{J}_{\Omega,W}\left(\theta(\mu)\right) \widehat{J}_T\left(\widehat{\theta}(\mu)\right)^{-1} \left(\partial\eta\left(\widehat{\theta}(\mu)\right)/\partial\theta\right)}}{\sqrt{T}},$$

and $\widetilde{\mathrm{CI}}(\mu) = [\tilde{c}_{\eta,T}(\alpha/2,\mu), \tilde{c}_{\eta,T}(1-\alpha/2,\mu)]$ for $\mu \in \Gamma$, then we have

$$\lim_{T\to\infty} \mathbb{P}_\mu^*\left\{\eta(\theta(\mu)) \in \widetilde{\mathrm{CI}}(\mu)\right\} = 1-\alpha, \tag{12}$$

and

$$\lim_{T\to\infty} \mathbb{P}^*\left\{\eta(\theta(\mu)) \in \cup_{\mu'\in\Gamma}\widetilde{\mathrm{CI}}(\mu'), \forall \mu \in \Gamma\right\} \geq 1-\alpha. \tag{13}$$

Proof. See Appendix 7.5. □

Theorem 4 verifies that intervals constructed making use of a normal approximation constructed conditional on fixed value of μ also have asymptotically correct coverage under the corresponding conditional measure. In practice, $\widehat{J}_T\left(\widehat{\theta}(\mu)\right)^{-1}$ in Theorem 4 can be computed by multiplying the variance-covariance matrix of the MCMC sequence $S = \left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(B)}\right)$, where B denotes the simulation sample size, by T in settings where MCMC is used to approximate the quasi-posterior at a fixed value of μ. As with Theorem 3, it is then immediate that a union of intervals obtained using this approach under a support restriction on μ delivers valid frequentist inference.

Note that we state these results for completeness and to verify that the quasi-Bayes approach can be used to deliver frequentist valid inference under only support restrictions as in Armstrong and Kolesár (2021). However, our chief interest is in using quasi-Bayes approaches in settings where informative prior information is of use. If valid frequentist inference under a support restriction is the goal, it is not clear there is much advantage to adopting the framework presented in this paper.

5. Conclusion

In this paper, we introduce Plausible GMM (PGMM), a quasi-Bayesian framework for inference in moment condition models that allows for potential misspecification. By placing a proper prior over the degree of misspecification, PGMM provides a flexible and transparent way to incorporate researchers' subjective beliefs about the plausibility of structural assumptions. This approach extends classical GMM by acknowledging that moment conditions are often credible but not exact, enabling more credible inference in the presence of model uncertainty.

Our theoretical contributions include posterior concentration results, new Bernstein-von Mises type approximations under partial identification, and decision-theoretic guarantees for quasi-Bayes procedures allowing diverging dimensions of parameters and moments. While not our main goal, we also provide an approach and results for using quasi-posteriors to obtain asymptotically valid frequentist inferential statements under support restrictions for the degree of misspecification.

Empirical applications illustrate the use of PGMM. In these examples, we see that PGMM intervals remain informative while allowing for subjective, but empirically motivated deviations away from dogmatic identifying assumptions. PGMM may thus offer a useful tool for applied researchers who wish to retain the structure of moment-based models while explicitly allowing for uncertainty about moments being perfectly satisfied.

6. APPENDIX: ADDITIONAL FIGURES

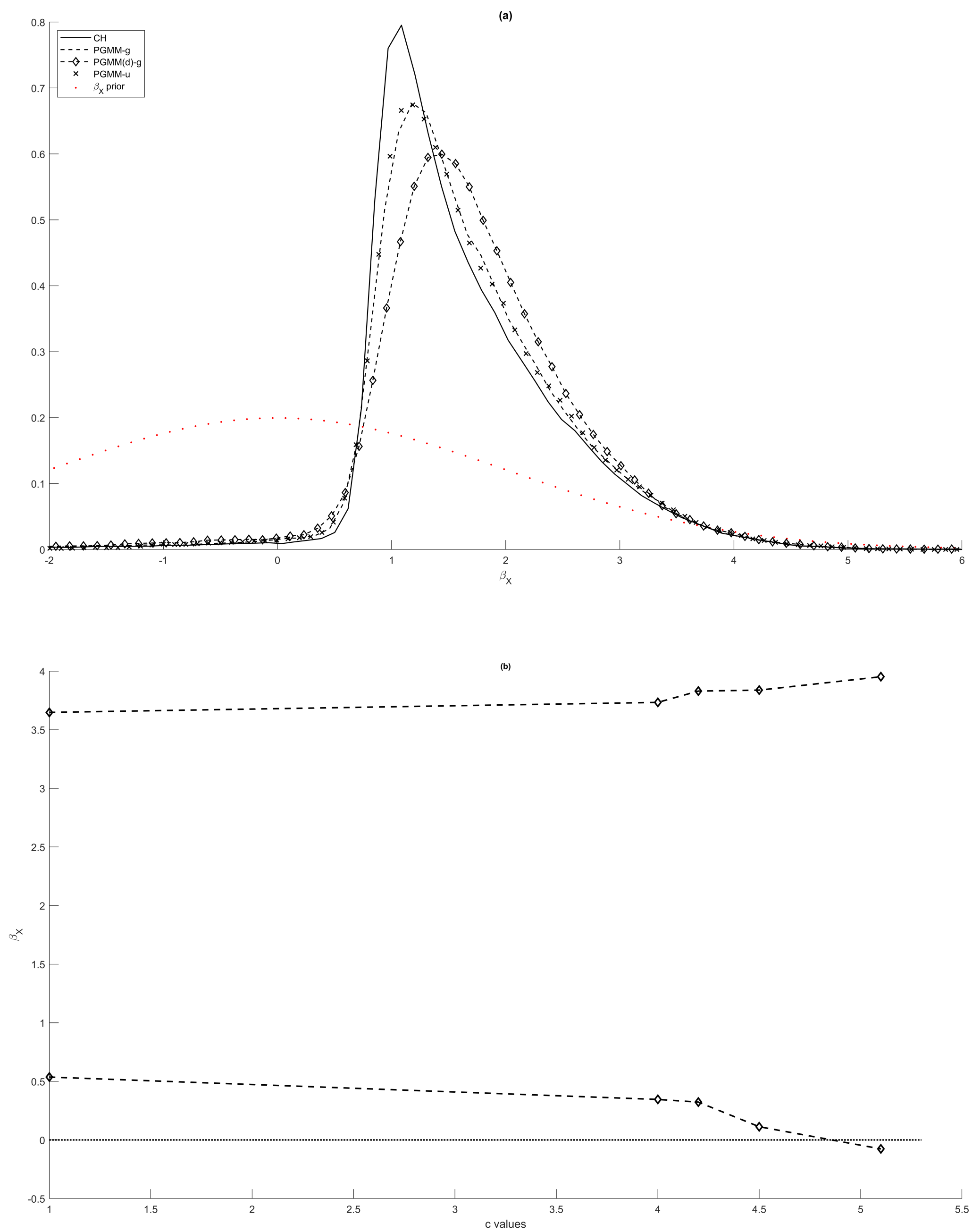

FIGURE 4. Upper panel (a): Linear IV (2) marginal (quasi-)posterior(s) for β_X and its marginal prior (red dotted curve); lower panel (b): Linear IV(2) 95% HPD intervals for β_X resulting from the PGMM approach along various priors for μ, i.e., $\mathcal{N}(0, c\Sigma_T\Omega_d\Sigma_T^\top)$ for various values of c.

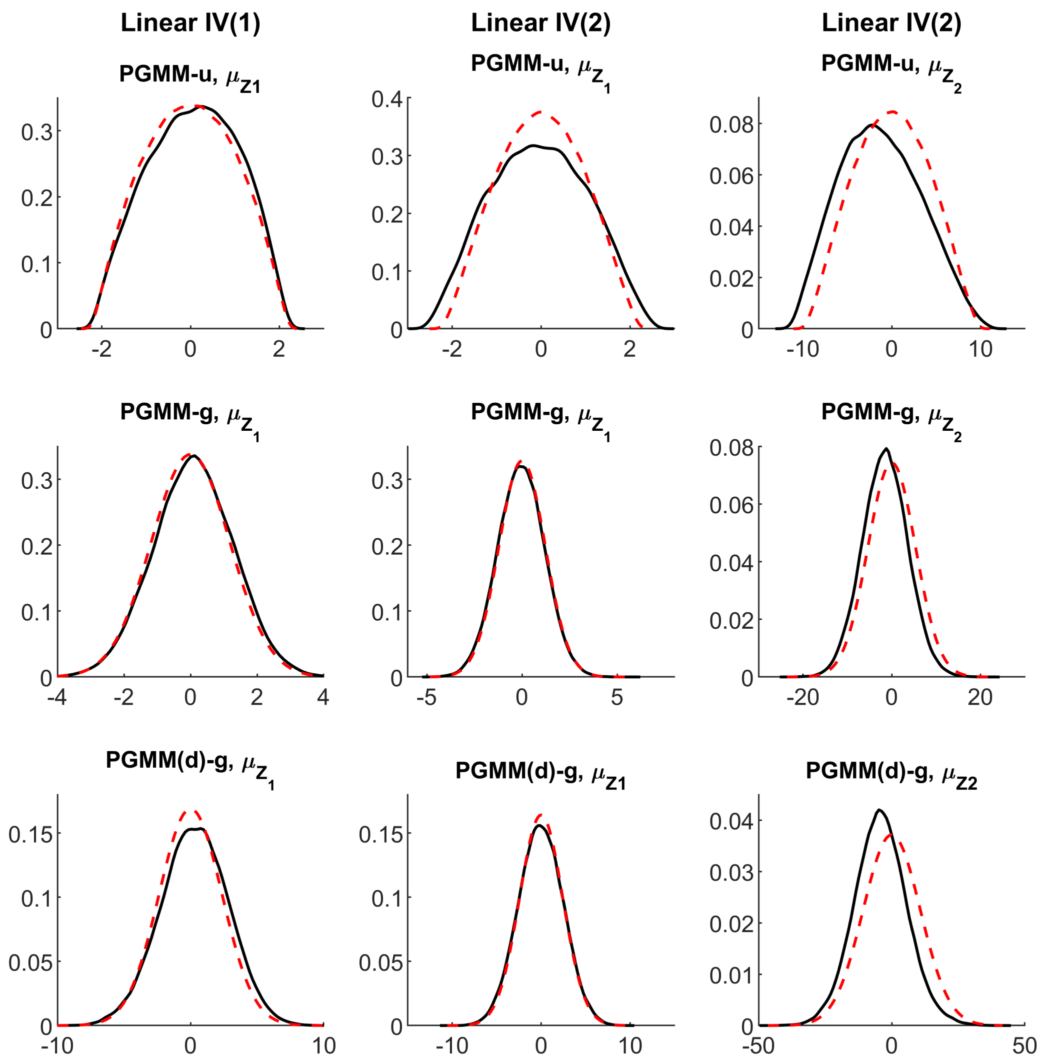

FIGURE 5. Marginal (quasi-)posteriors and priors for selected μ's (entries corresponding to moments constructed with IVs) in Linear IV(1) and Linear IV(2).

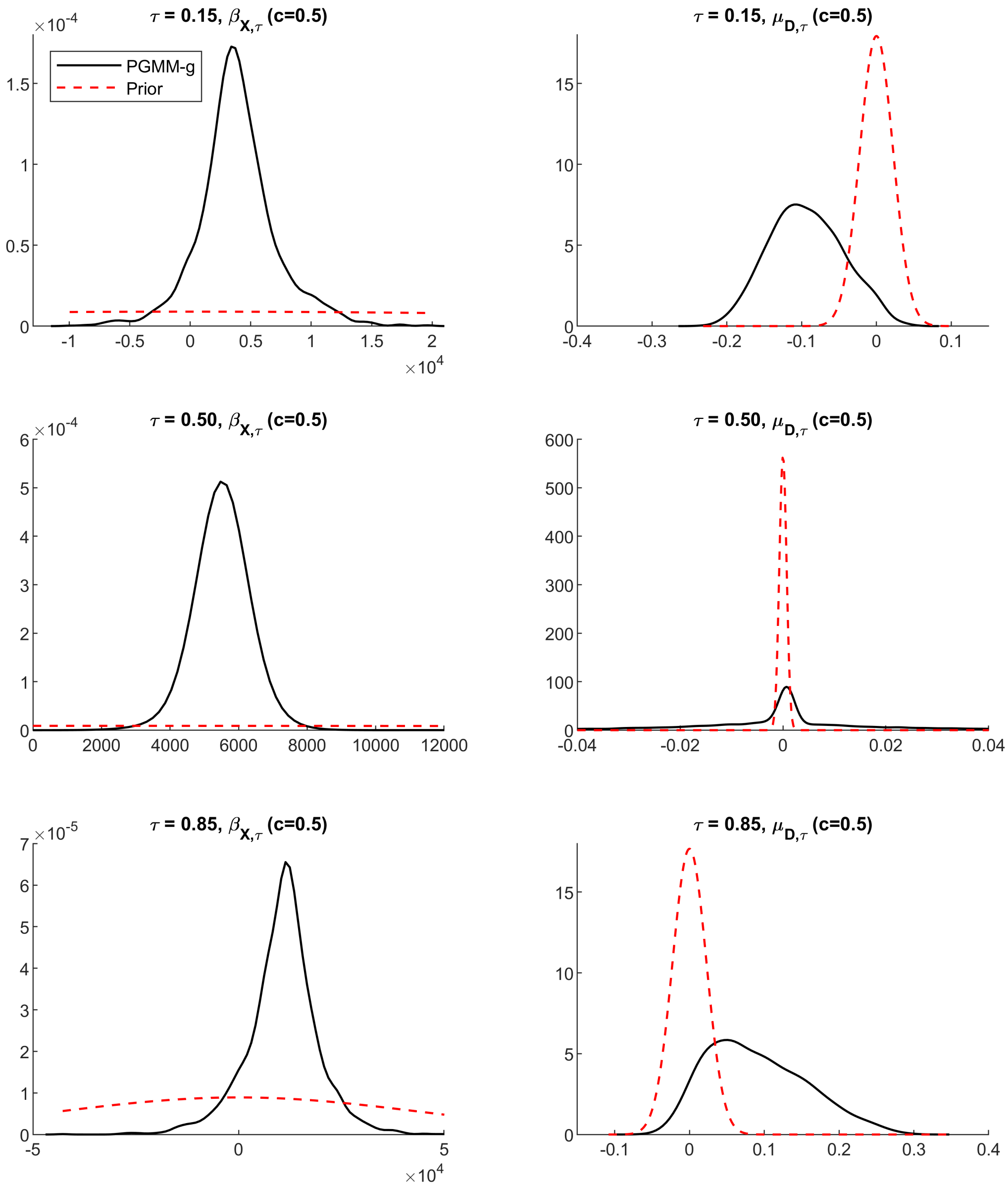

FIGURE 6. PGMM marginal quasi-posteriors (solid black curves) of $\beta_{X,\tau}$ for the Gaussian prior over μ across various values of τ and $c = 0.5$. The red dashed curves represent the marginal prior density curves for these parameters.

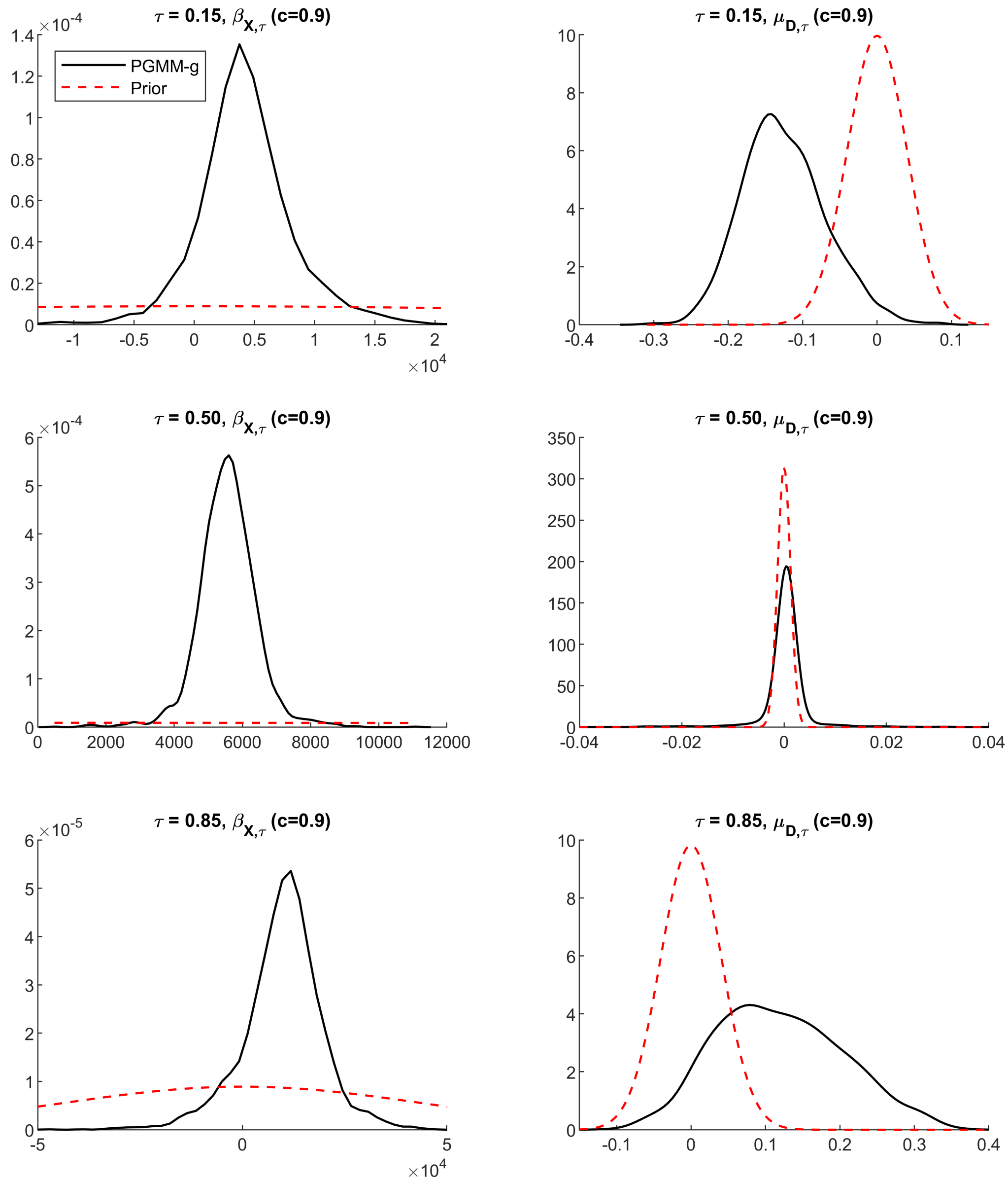

FIGURE 7. PGMM marginal quasi-posteriors (solid black curves) of $\beta_{X,\tau}$ for the Gaussian prior over μ across various values of τ and $c = 0.9$. The red dashed curves represent the marginal prior density curves for these parameters.

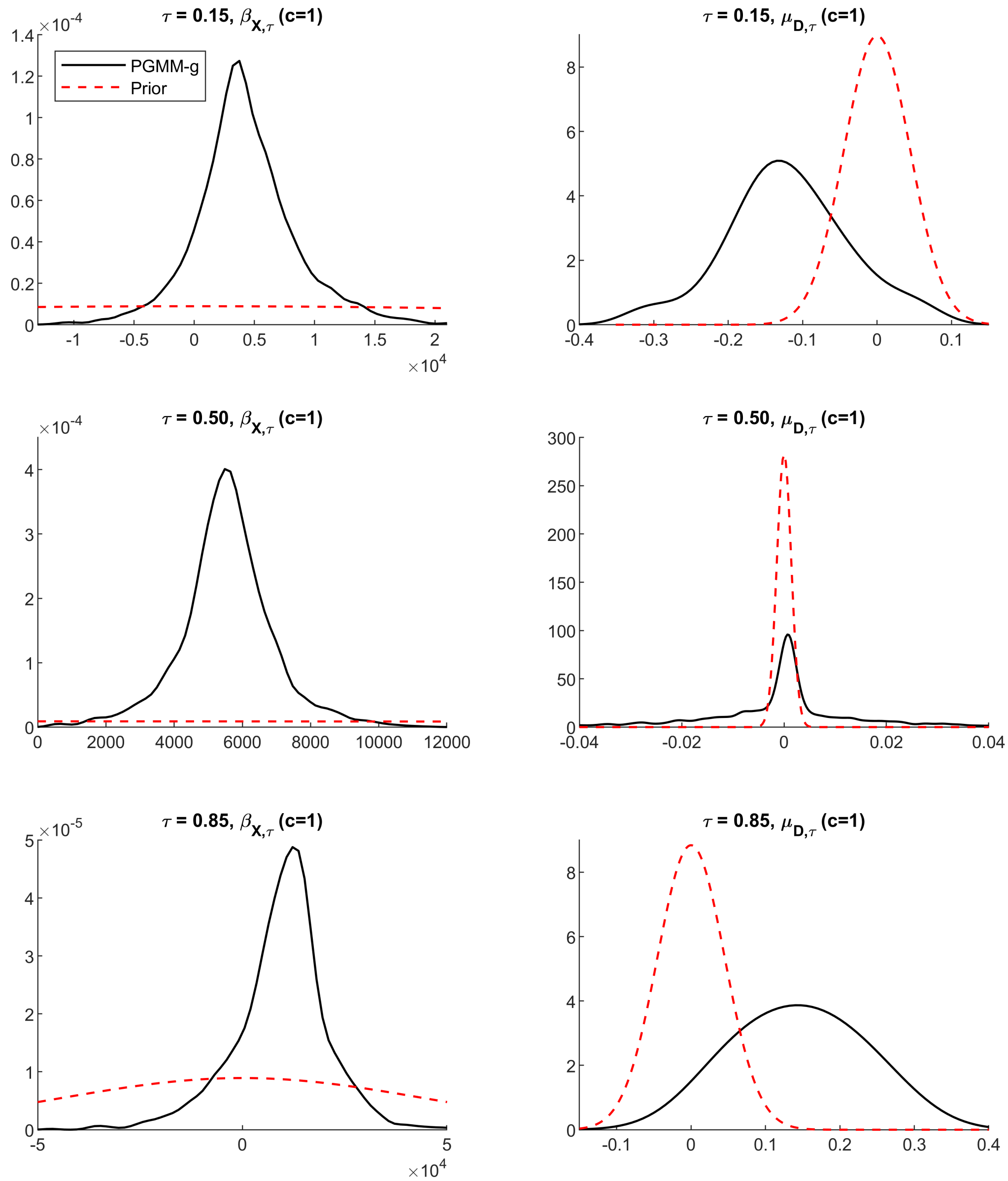

FIGURE 8. PGMM marginal quasi-posteriors (solid black curves) of $\beta_{X,\tau}$ for the Gaussian prior over μ across various values of τ and $c = 1$. The red dashed curves represent the marginal prior density curves for these parameters.

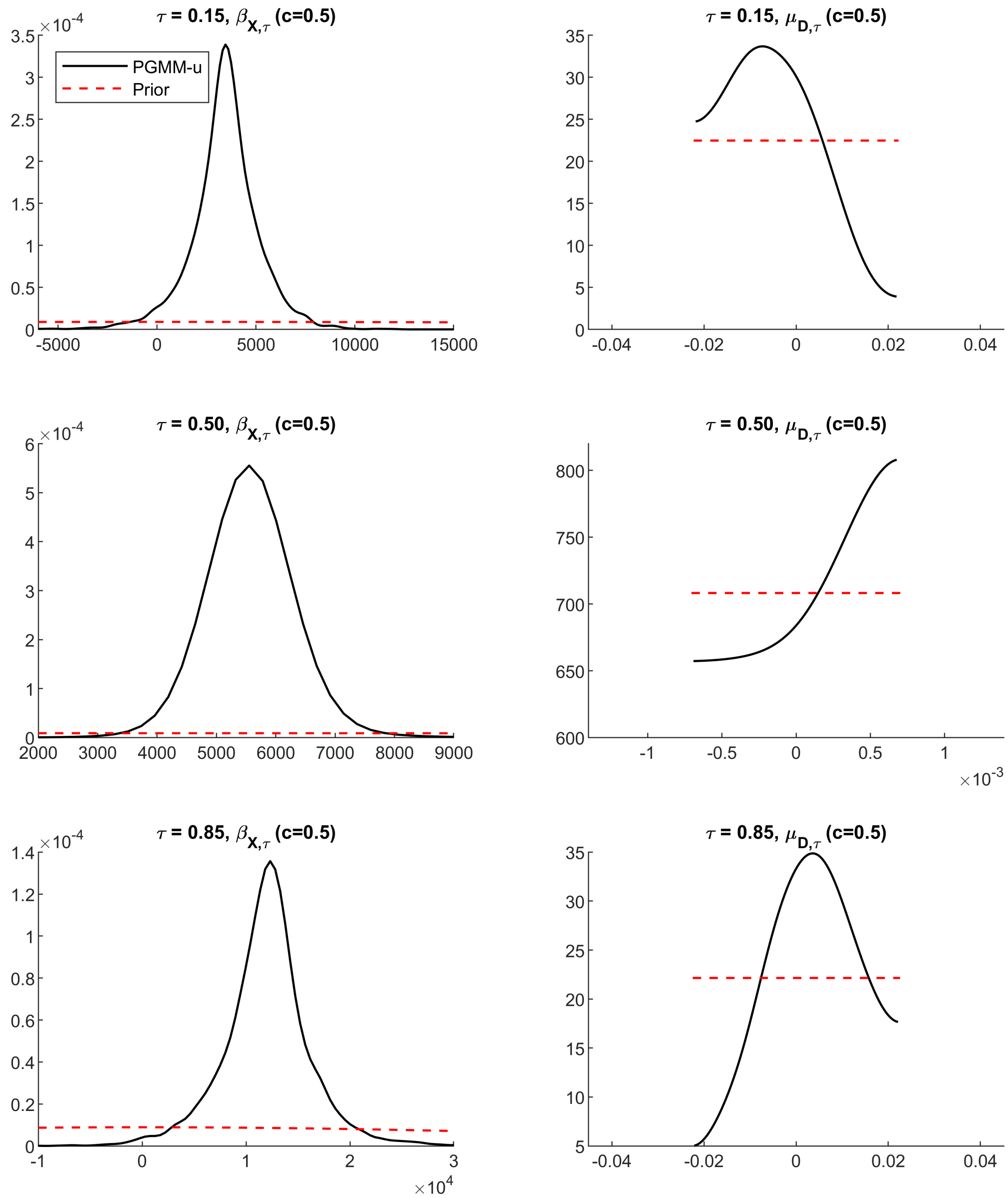

FIGURE 9. PGMM marginal quasi-posteriors (solid black curves) of $\beta_{X,\tau}$ for the uniform prior over μ across various values of τ and $c = 0.5$. The red dashed curves represent the marginal prior density curves for these parameters.

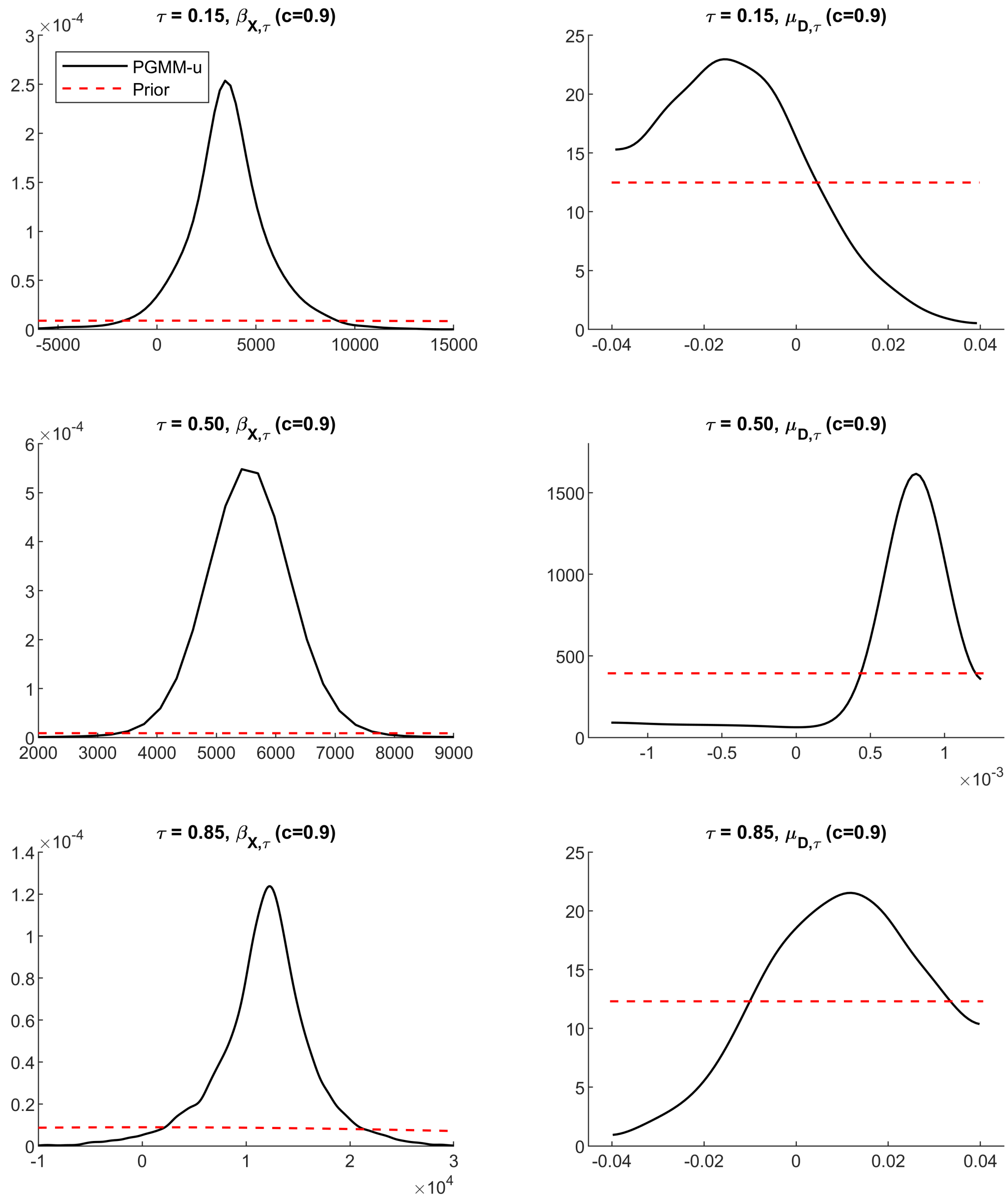

FIGURE 10. PGMM marginal quasi-posteriors (solid black curves) of $\beta_{X,\tau}$ for the uniform prior over μ across various values of τ and $c = 0.9$. The red dashed curves represent the marginal prior density curves for these parameters.

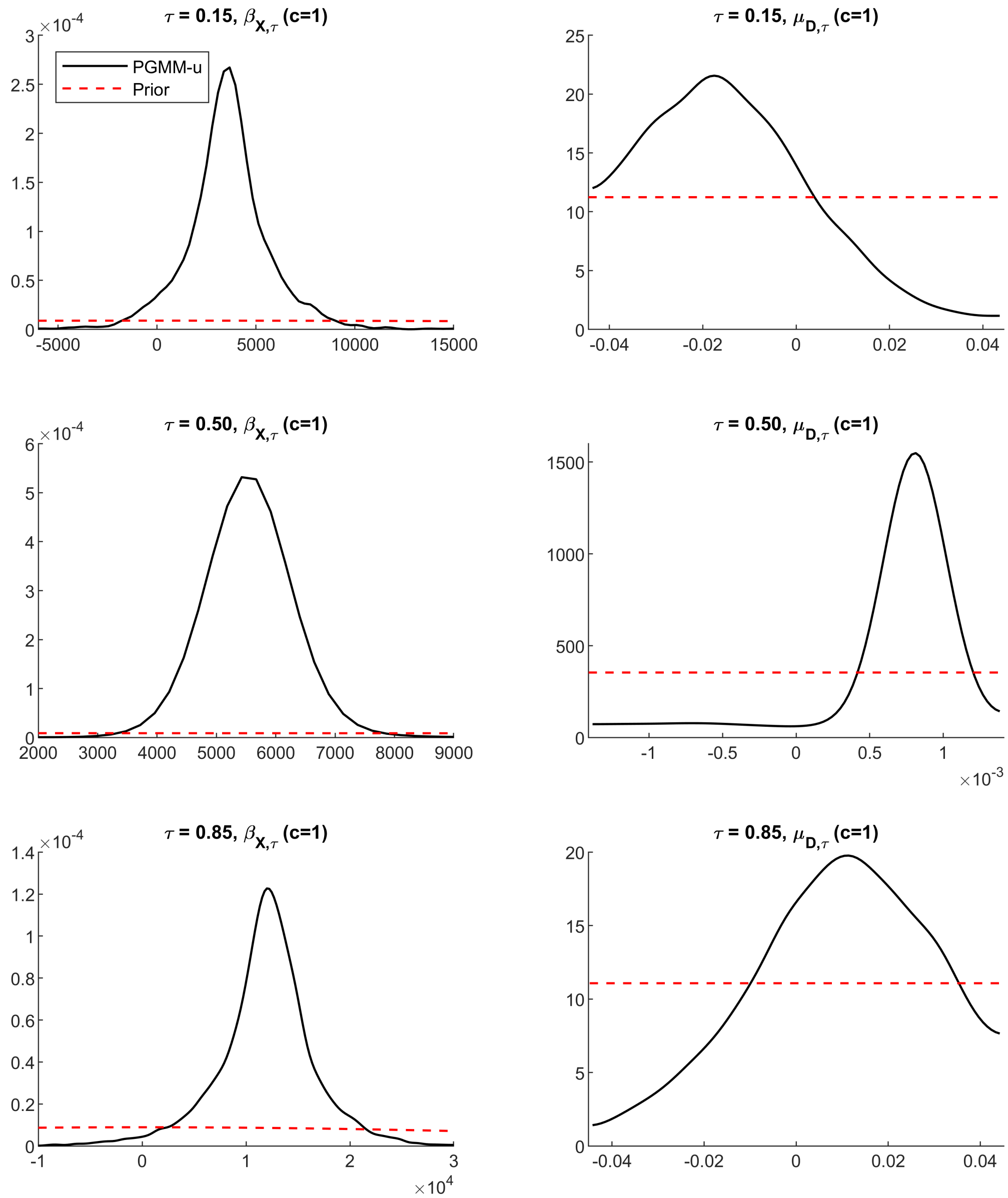

FIGURE 11. PGMM marginal quasi-posteriors (solid black curves) of $\beta_{X,\tau}$ for the uniform prior over μ across various values of τ and $c = 1$. The red dashed curves represent the marginal prior density curves for these parameters.

7. Appendix: Proofs

Notation We denote 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 Frobenious norm by $\|A\|_F$. The trace of an $n \times n$ square matrix A is defined as $tr(A) = \sum_i a_{i,i}$ and its determinant is denoted by $\det(A)$. For positive semi-definite matrices A, B, we write $A \geq B$ if $A - B$ is positive semi-definite. "w.p.a.1." to denote "with probability approaching one". For $s > 0$ and a random vector X, we say $X \in \mathscr{L}^s$ if $\|X\|_s = [\mathbb{E}(|X|^s)]^{1/s} < \infty$.

7.1. **Proof of Theorem 1.**

$$p_T(\theta, \mu) \propto \pi(\theta) \exp(-\frac{T}{2} \|\widehat{m}(\theta) - \mu\|^2_{\widehat{\Omega}_T(\theta)^{-1}}) \exp(-(T\mu^\top \Lambda^{-1} \mu)/2).$$

We shall now prove that it is indeed true that,

$$p_T(\theta) \propto \exp(-\frac{T}{2} \|\widehat{m}(\theta) - \mu_0\|^2_{\widehat{A}_{T,\theta}}).$$

Without loss of generality, we prove for $\mu_0 = 0$. Let $\mu \sim N(\mu_0, T^{-1}\Lambda)$. Let

$$C_w^2 = T(\Lambda^{-1} + \widehat{\Omega}_T(\theta)^{-1}),$$

$$C_w C_g = T\widehat{\Omega}_T(\theta)^{-1} \widehat{m}(\theta).$$

Then we have the following,

$$\begin{aligned}
p_T(\theta, \mu) &\propto \pi(\theta) \exp(-\frac{1}{2} \|\widehat{m}(\theta) - \mu\|^2_{T\widehat{\Omega}_T(\theta)^{-1}} - \|\mu\|^2_{T\Lambda^{-1}}/2) \\
&\propto \pi(\theta) \exp(-\frac{1}{2} \|\widehat{m}(\theta)\|^2_{T\widehat{\Omega}_T(\theta)^{-1}} + T\mu^\top \widehat{\Omega}_T(\theta)^{-1} \widehat{m}(\theta) - \|\mu\|^2_{T\widehat{\Omega}_T(\theta)^{-1}}/2 - \|\mu\|_{T\Lambda^{-1}}/2) \\
&\propto \pi(\theta) \exp(-\frac{1}{2} (\|\widehat{m}(\theta)\|^2_{T\widehat{\Omega}_T(\theta)^{-1}}) \exp(+T\mu^\top \widehat{\Omega}_T(\theta)^{-1} \widehat{m}(\theta) - \|\mu\|^2_{(\Lambda^{-1} + \widehat{\Omega}_T(\theta)^{-1})T}/2) \\
&\propto \pi(\theta) \exp(-\frac{1}{2} (\|\widehat{m}(\theta)\|^2_{T\widehat{\Omega}_T(\theta)^{-1}}) \exp(+T\mu^\top \widehat{\Omega}_T(\theta)^{-1} \widehat{m}(\theta) - \|\mu\|^2_{(\Lambda^{-1} + \widehat{\Omega}_T(\theta)^{-1})T}/2) \\
&\propto \pi(\theta) \exp(-\frac{1}{2} (\|\widehat{m}(\theta)\|^2_{T\widehat{\Omega}_T(\theta)^{-1}}) \exp(+\mu^\top C_w C_g - \|\mu\|^2_{C_w^2}/2 - C_g^\top C_g/2) \exp(+C_g^\top C_g/2) \\
&\propto \pi(\theta) \exp(-\frac{1}{2} (\|\widehat{m}(\theta)\|^2_{T\widehat{\Omega}_T(\theta)^{-1}}) \exp(-(C_w\mu - C_g)^\top (C_w\mu - C_g)/2) \exp(C_g^\top C_g/2) \\
&\propto \pi(\theta) \exp(-\frac{1}{2} (\|\widehat{m}(\theta)\|^2_{T\widehat{\Omega}_T(\theta)^{-1}}) \exp(-(\mu - C_w^{-1} C_g)^\top C_w^2 (\mu - C_w^{-1} C_g)/2) \exp(C_g^\top C_g/2).
\end{aligned}$$

Plugging in the definition of C_g, we have that

$$\begin{aligned}
\int p_T(\theta, \mu) d\mu &\propto \pi(\theta) \exp(-\frac{1}{2} (\|\widehat{m}(\theta)\|^2_{T\widehat{\Omega}_T(\theta)^{-1}})) \exp(T^2 \widehat{m}(\theta)^\top \widehat{\Omega}_T(\theta)^{-1} C_w^{-2} \widehat{\Omega}_T(\theta)^{-1} \widehat{m}(\theta)/2) \\
&\quad * \sqrt{2\pi \det(C_w^2)} \\
&\propto \pi(\theta) \exp(-\frac{1}{2} (\|\widehat{m}(\theta)\|^2_{T(\widehat{\Omega}_T(\theta)^{-1} - \widehat{\Omega}_T(\theta)^{-1} T C_w^{-2} \widehat{\Omega}_T(\theta)^{-1})})).
\end{aligned}$$

Then, the results directly follow the proof of Theorem 1 in Chernozhukov and Hong (2003) for fixed Λ. In the scenario where $c \ll \Lambda \ll T$ (c is a positive constant), we need to replace M in the proof of Theorem 1 in Chernozhukov and Hong (2003) by a slowly increasing M_T, e.g., $M_T = M\|A_{\theta(\mu_0)}\|^{-\frac{1}{2}}$, and the remaining proofs proceed with a similar rationale.

7.2. **Proof of Theorems 2 and other intermediate results.** It shall be noted that Assumption 3 imply that

$$\mathrm{tr}\left(G(\theta)^\top W(\theta)G(\theta)\right) \lesssim k,$$

and

$$\mathrm{tr}(TG(\theta)^\top(\widehat{m}(\theta)-\mu)^\top W(\theta)(\widehat{m}(\theta)-\mu)G(\theta)) = O_p(k),$$
$$0 < \lambda_{\min}(W(\theta)G(\theta)G(\theta)^\top W(\theta)) < \lambda_{\max}(W(\theta)G(\theta)G(\theta)^\top W(\theta)) \lesssim 1.$$

Define

$$C_{m,\mu} = G^\top(\theta(\mu))W(\theta(\mu))(\widehat{m}(\theta(\mu))-\mu),$$
$$C_{w,\mu} = G^\top(\theta(\mu))W(\theta(\mu))G(\theta(\mu))$$

and

$$2C_\mu = -2\log\pi(\mu) - 2\log(\pi(\theta(\mu))) + T(\widehat{m}(\theta(\mu))-\mu)^\top W(\theta(\mu))(\widehat{m}(\theta(\mu))-\mu).$$

Then let us analyse $V_T(h(.),\theta,\mu)$,

$$\begin{aligned}
& V_T(h(.),\theta,\mu) - 2\log\pi(\mu) \\
= \quad & 2T[(\theta-\theta(\mu))^\top G^\top(\theta(\mu))]W(\theta(\mu))(\widehat{m}(\theta(\mu))-\mu) \\
& +T[(\theta-\theta(\mu))^\top G^\top(\theta(\mu))]W(\theta(\mu))[G(\theta(\mu))(\theta-\theta(\mu))] \\
& -2\log\pi(\mu) - 2\log(\pi(\theta(\mu))) + T(\widehat{m}(\theta(\mu))-\mu)^\top W(\theta(\mu))(\widehat{m}(\theta(\mu))-\mu) \\
= \quad & 2T(\theta-\theta(\mu))^\top C_{m,\mu} + T(\theta-\theta(\mu))^\top C_{w,\mu}(\theta-\theta(\mu)) + 2C_\mu.
\end{aligned}$$

We know that conditioning on μ, $\exp(-\frac{V_T(h(.),\theta,\mu)-2\log\pi(\mu)}{2})$ is proportional to the log-likelihood of the density function of $N(-C_{w,\mu}^{-1}C_{m,\mu}, (TC_{w,\mu})^{-1})$.

By Assumption 6,

$$\begin{aligned}
tr(C_{m,\mu}C_{m,\mu}^\top) &= tr((\widehat{m}(\theta(\mu))-\mu)^\top W(\theta(\mu))G(\theta(\mu))G^\top(\theta(\mu))W(\theta(\mu))(\widehat{m}(\theta(\mu))-\mu)) \\
&= O_p(tr(G^\top(\theta(\mu))W(\theta(\mu))\mathbb{E}[(\widehat{m}(\theta(\mu))-\mu)(\widehat{m}(\theta(\mu))-\mu)^\top]W(\theta(\mu))G(\theta(\mu)))) \\
&= O_p(\frac{k}{T}).
\end{aligned}$$

By Assumption 6, and $\lambda_{\max}(G(\theta(\mu))G^\top(\theta(\mu))) = \lambda_{\max}(G^\top(\theta(\mu))G(\theta(\mu))) \lesssim 1$.

To derive the conclusion, note that $\int_{\mathcal{M}}\int_\Theta \left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\left|p_T(\theta,\mu)-N_T(\theta,\mu)\right| \mathrm{d}\theta d\mu$ may divide into the following parts:

$$\int_{\mathcal{M}}\int_\Theta \left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\left|p_T(\theta,\mu)-N_T(\theta,\mu)\right| \mathrm{d}\theta d\mu = \mathscr{R}_{1,T} + \mathscr{R}_{2,T}.$$

where we denote $\mathscr{R}_{1,T} = \int_{B_\varepsilon}\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\left|p_T(\theta,\mu)-N_T(\theta,\mu)\right|\mathrm{d}\theta d\mu$ and $\mathscr{R}_{2,T} = \int_{B_\varepsilon^c}\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\left|p_T(\theta,\mu)-N_T(\theta,\mu)\right|\mathrm{d}\theta d\mu$.

To look at $\mathscr{R}_{1,T}$, we first look at $\int_{B_\varepsilon}\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)N_T(\theta,\mu)\left|p_T(\theta,\mu)/N_T(\theta,\mu)-1\right|\mathrm{d}\theta d\mu$.

Let the integral ratio be

$$c^* = \frac{\int_\Xi \exp(-V_T(h(.),\theta,\mu)/2+\log(\pi(\mu)))d\theta d\mu}{\int_\Xi \exp(\frac{1}{2}Q_T(\theta,\mu)+\log\pi(\mu,\theta))d\theta d\mu}.$$

By Lemma 3, under Assumptions 7-8, we have $|c^*|\lesssim 1$.

We define $N_T(\theta|\mu)\propto \frac{N_T(\theta,\mu)}{\int_\Theta N_T(\theta,\mu)d\theta}$ and $N_T(\mu)\propto\int_\Theta N_T(\theta,\mu)d\theta$ for $\mu\in\mathscr{M}$. It is not hard to see that condition on μ, the density function $N_T(\theta|\mu)$ is proportional to a density function of a multivariate Gaussian random variable with mean $-C_{w,\mu}^{-1}C_{m,\mu}$ and variance $T^{-1}C_{w,\mu}^{-1}$, following the fact about Gaussian Integral such that, with $\boldsymbol{x}\in\mathbb{R}^n$, we have

$$\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top C_{w,\mu}\boldsymbol{x}+{C_{m,\mu}}^\top\boldsymbol{x}\right)dx_1dx_2\dots dx_n = \frac{(2\pi)^{n/2}}{\det(C_{w,\mu})^{1/2}}\exp\left[\frac{1}{2}{C_{m,\mu}}^\top {C_{w,\mu}}^{-1}C_{m,\mu}\right].$$

And we define $\mathbb{E}_{N_T(\theta|\mu)}(.)$ as taking expectation under the measure corresponding to the density function $N_T(\theta|\mu)$. We denote $\mathbb{E}_\gamma(.)$ as taking expectation under a standard multivariate Gaussian distribution with an identity variance covariance matrix. And $\mathbb{P}_\gamma(.)$ is the probability corresponding to γ.

Note that, $\frac{p_T(\theta,\mu)}{N_T(\theta,\mu)} = \frac{\exp\left(\frac{1}{2}Q_T(\theta,\mu)+V_T(h(.),\theta,\mu)/2+\log(\pi(\theta))\right)}{c^*} = \frac{\exp(R_T(\theta,\mu))}{c^*}$, let $a(\theta,\mu)=\exp(R_T(\theta,\mu))/c^*-1$, and then for a positive constant $\delta>0$,

$$\begin{aligned}
\mathscr{R}_{1,T} &= \int_{B_\varepsilon}\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)N_T(\theta,\mu)\left|\frac{\exp(R_T(\theta,\mu))}{c^*}-1\right|\mathrm{d}\theta d\mu\\
&= \int_{B_\varepsilon}\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)N_T(\theta,\mu)\left|a(\theta,\mu)\right|\mathrm{d}\theta d\mu\\
&= \int_{B_\varepsilon}\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)N_T(\theta|\mu)\left|a(\theta,\mu)\right|\mathrm{d}\theta N_T(\mu) d\mu\\
&\le \frac{\delta(\log T)^3\sqrt{k^3}}{\sqrt{T}}\int_{B_\varepsilon}N_T(\theta|\mu)\left|\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\right|\mathrm{d}\theta N_T(\mu)d\mu\\
&\le \frac{\delta(\log T)^3\sqrt{k^3}}{\sqrt{T}}\\
&\quad \int_{\Gamma_\epsilon}\mathbb{E}_\gamma\left(\left\|\sqrt{T}^{-1}C_{w,\mu}^{-1/2}(\gamma-C_{w,\mu}^{-1/2}C_{m,\mu})\right\|^\kappa \mathbf{1}(\left\|\sqrt{T}^{-1}C_{w,\mu}^{-1/2}(\gamma-C_{w,\mu}^{-1/2}C_{m,\mu})\right\|\le\varepsilon/\sqrt{T})\middle|\mu\right)N_T(\mu)d\mu\\
&\lesssim \frac{(\log T)^3\delta(\sqrt{k})^{\kappa+3}}{\sqrt{T}},
\end{aligned}$$

where we let $\Gamma_\epsilon=\left\{\mu\in\mathscr{M}:\|\mu-\nu(\mu)\|\le\log(T)\epsilon\right\}$ and the first inequality is due to Lemma 3, Assumptions 8, and the third inequality is due to Lemma 2 and the property of Gaussian density.

Denote $\mathbb{P}_{N_T(\theta|\mu)}(\cdot)$ as the conditional probability measure function of $\theta-\theta(\nu(\mu))$ conditioning on a fixed value of μ corresponding to the density $N_T(\theta|\mu)$.

For $\mathscr{R}_{1,T}$, we essentially only need to look at the region $\bar{B}^c_\varepsilon=\{(\theta,\mu):\sqrt{T}\|h(\theta,\mu)\|>\varepsilon,\theta\in\Theta,\mu\in\mathscr{M}\}$ as the integration over the remaining region is an $o_p(1)$ term (see also, e.g., Theorem 5 from Andrews and Mikusheva (2022)). On $\bar{B}^c_\varepsilon$, Assumption 8 implies that $\left|\exp(R_T(\theta,\mu))/c^*-1\right|\le \exp(-C_0\|\theta-\theta(\nu(\mu))\|_{C_{w,\mu}}\varepsilon\sqrt{T}+C_0\varepsilon^2/2+T\|\theta-\theta(\nu(\mu))\|^2_{C_{w,\mu}}/2)/c^*+1$.

We define $B^c_{\varepsilon,\theta}=\{\theta:\|(\theta-\theta(\upsilon(\mu)))\|>\sqrt{T}^{-1}\|G(\theta(\upsilon(\mu)))\|^{-1}\varepsilon,\mu\in\mathscr{M}\}$, so that in this set $\|C^{-1/2}_{w,\mu}(\sqrt{T}^{-1}\gamma-C^{-1/2}_{w,\mu}C_{m,\mu})\|>c\sqrt{T}^{-1}\sqrt{k}\log T$ for a positive c. Additionally, we define $B_\gamma=\{\gamma\in\mathbb{R}^k:\|\gamma\|>\sqrt{T}^{-1}\left(c\sqrt{k\log T}-\|C^{-1/2}_{w,\mu}C_{m,\mu}\|\right)\|C^{-1/2}_{w,\mu}\sqrt{T}^{-1}\|^{-1}\}$, and let $\mathscr{B}_T(\theta-\theta(\nu(\mu))=-C_0\|\theta-\theta(\nu(\mu))\|\varepsilon\sqrt{T}+C_0\varepsilon^2/2+T\|\theta-\theta(\nu(\mu))\|^2_{C_{w,\mu}}/2$. Thus it implies that γ satisfying $\|C^{-1/2}_{w,\mu}(\sqrt{T}^{-1}\gamma-C^{-1/2}_{w,\mu}C_{m,\mu})\|>c\sqrt{T}^{-1}\sqrt{k}\log T$ is contained in the ball B_γ.

To handle $\mathscr{R}_{2,T}$, first we define a $\mathscr{R}_{22,T}$ term as,

$$
\begin{aligned}
\mathscr{R}_{22,T}&=\int_{\bar{B}^c_\varepsilon}\exp(\mathscr{B}_T(\theta-\theta(\nu(\mu))))N_T(\theta|\mu)\left|\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\right|\mathrm{d}\theta N_T(\mu)d\mu\\
&\lesssim\int_{\mathscr{M}}\mathbb{E}_{N_T(\theta|\mu)}[\exp(\mathscr{B}_T(\theta-\theta(\nu(\mu)))\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\mathbf{1}(\theta\in B^c_{\varepsilon,\theta})]N_T(\mu)d\mu\\
&\lesssim\int_{\mathscr{M}}\|C^{-1/2}_{w,\mu}\sqrt{T}^{-1}\|\mathbb{E}_\gamma[\exp(\mathscr{B}_T(C^{-1/2}_{w,\mu}[\sqrt{T}^{-1}\gamma-C^{-1/2}_{w,\mu}C_{m,\mu}])\\
&\left(1+\|C^{-1/2}_{w,\mu}(\sqrt{T}^{-1}\gamma-C^{-1/2}_{w,\mu}C_{m,\mu})\|^\kappa\right)\mathbf{1}(\gamma\in B_\gamma)]N_T(\mu)d\mu\\
&\le\int_{\mathscr{M}}\|C^{-1/2}_{w,\mu}\sqrt{T}^{-1}\|\mathbb{E}_\gamma[\exp(\mathscr{B}_T(C^{-1/2}_{w,\mu}\sqrt{T}^{-1}\gamma)+\|C^{-1}_{w,\mu}C_{m,\mu}\|C_0\varepsilon\sqrt{T}/2\\
&+T\|C^{-1}_{w,\mu}C_{m,\mu}\|^2_{C_{w,\mu}}/2)\left(1+\|C^{-1/2}_{w,\mu}\sqrt{T}^{-1}\gamma\|^\kappa+\|C^{-1}_{w,\mu}C_{m,\mu}\|^\kappa\right)]\mathbf{1}(\gamma\in B_\gamma)N_T(\mu)d\mu\\
&\le C\sup_{\mu\in\mathscr{M}}\|C^{-1/2}_{w,\mu}\sqrt{T}^{-1}\|\sqrt{k}\exp(-(\varepsilon\vee\sqrt{k})),
\end{aligned}
$$

where the bound is by lemma 10 as Gaussian in integral lemma and implied by our Assumptions 6 and 8 .

$$
\begin{aligned}
\mathscr{R}_{2,T} &= \int_{\bar{B}_\varepsilon^c} \left(1 + \|\theta - \theta(\nu(\mu))\|^\kappa\right) N_T(\theta,\mu) \left|\exp(R_T(\theta,\mu))/c^* - 1\right| \mathrm{d}\theta d\mu \\
&\lesssim \int_{\bar{B}_\varepsilon^c} N_T(\theta|\mu) \left|\left(1 + \|\theta - \theta(\nu(\mu))\|^\kappa\right)\right| \mathrm{d}\theta N_T(\mu) d\mu + \mathscr{R}_{22,T} \\
&\leq \int_0^\infty \int_{\mathcal{M}} \mathbb{P}_{N_T(\theta|\mu)}(\|\theta - \theta(\nu(\mu))\|^\kappa > \varepsilon/\sqrt{T} \vee x|\mu) N_T(\mu) d\mu dx + \mathscr{R}_{22,T} \\
&\leq \int_0^\infty \int_{\mathcal{M}} \mathbb{P}_\gamma(\|C_{w,\mu}^{-1/2}(\sqrt{T}^{-1}\gamma - C_{w,\mu}^{-1/2} C_{m,\mu})\|^\kappa > \varepsilon/\sqrt{T} \vee x|\mu) N_T(\mu) d\mu dx + \mathscr{R}_{22,T} \\
&\leq \sqrt{k}^\kappa \int_0^\infty \int_{\mathcal{M}} \exp(-(\varepsilon \vee (x - \sqrt{k}))) N_T(\mu) d\mu dx \to 0,
\end{aligned}
$$

where $\mathbb{P}_\gamma(.)$ denotes probability corresponding to an k-dimensional standard Gaussian distribution, the first inequality is due to Assumption 8, the second inequality is due to Lemma 4 and the last inequality is due to Lemma 2.

7.3. **Proof of Lemma 1.**

Proof. We aim to establish that

$$
\int_\mu \mathbb{P}_\mu\left(\theta(\mu) \in PR_T\right) \pi(\mu) d\mu = 1 - \alpha + o_p(1). \tag{14}
$$

We begin by considering the case under the conditions of Theorem 1, as the argument for Theorem 2 proceeds analogously.

We know from Theorem 1 that $PR_T(\alpha) = \left\{\theta : T(\widehat{m}(\theta))^\top A_{\theta_0}(\widehat{m}(\theta)) \leq Z_\alpha\right\}$ such that Z_α satisfies the following condition

$$
\mathbb{P}_{T,\theta}\left(T(\widehat{m}(\theta)^\top A_{\theta_0}\widehat{m}(\theta)) \leq Z_\alpha\right) = 1 - \alpha,
$$

where $\mathbb{P}_{T,\theta}$ denotes the probability measure of $\widehat{m}(\theta) - \mu + \mu$ corresponding to the marginal posterior $p_T(\theta)$, which in this case is Gaussian with mean 0 and covariance matrix $A_{\theta_0}^{-1}$. Therefore, Z_α is the $(1-\alpha)$ quantile of a chi-squared distribution with q degrees of freedom. To verify equation (14), it suffices to show that

$$
\int \mathbb{P}_\mu\left(T(\widehat{m}(\theta(\mu)) - \mu + \mu)^\top A_{\theta_0}(\widehat{m}(\theta(\mu)) - \mu + \mu)) \leq Z_\alpha\right) f_\mu(\mu) d\mu = 1 - \alpha + o_p(1), \tag{15}
$$

where $A_{\theta_0} = \Lambda^{-1} - \Lambda^{-1}\left(\Omega(\theta_0)^{-1} + \Lambda^{-1}\right)^{-1}\Lambda^{-1}$, so that $A_{\theta_0}^{-1} = \Omega(\theta_0) + \Lambda$ by the Woodbury matrix identity.

Note that under $\mathbb{P}$, $m(\theta(\mu)) = \mu$ following the distribution specified by $f_\mu(\cdot)$ corresponding to the local Gaussian prior, while for a given μ, under $\mathbb{P}_\mu$, $\widehat{m}(\theta(\mu)) - \mu$ is asymptotically Gaussian with mean zero and variance $T^{-1}\Omega(\theta_0)$. Consequently, the sum of $\widehat{m}(\theta(\mu)) - \mu$ and μ is Gaussian with mean zero and variance $A_{\theta_0}^{-1}$ under $\mathbb{P}$. Thus by the definition of the quantile Z_α, the following is true:

$$\int \mathbb{P}_\mu \left(T(\widehat{m}(\theta(\mu)) - \mu + \mu))^\top A_{\theta_0}(\widehat{m}(\theta(\mu)) - \mu + \mu) \le Z_\alpha \right) f_\mu(\mu) d\mu = 1 - \alpha,$$

which then implies equation (15) upon noting that $\|\mu - \mu_0\| = O_p(\frac{\sqrt{q}}{\sqrt{T}})$ in this local misspecficiation case.

We now generalize the preceding argument to the nonlocal case under the conditions of Theorem 2. Here, we have that by definition $PR_T(\alpha) = \left\{\theta : \int_\mu p_T(\theta, \mu) d\mu \ge Z_\alpha \right\} = \left\{\theta : p_T(\theta) \ge Z_\alpha \right\}$ where Z_α is chosen to satisfy

$$\mathbb{P}_{T,\theta} \left(p_T(\theta) \ge Z_\alpha \right) = 1 - \alpha,$$

and $\mathbb{P}_{T,\theta}$ corresponds to the marginal posterior $p_T(\theta)$. With a slight abuse of notation, we treat $p_T(\theta)$ as a function of the sample moments $\widehat{m}(\theta)$, rather than of θ directly. Define this function as $\tilde{p}_T(\widehat{m}(\theta)) := p_T(\theta)$. Then the above condition can be rewritten as

$$\mathbb{P}_{T,\theta} \left(\tilde{p}_T(\widehat{m}(\theta)) \ge Z_\alpha \right) = 1 - \alpha. \tag{16}$$

To verify equation (14), it suffices to show that

$$\int \mathbb{P}_\mu \left(\tilde{p}_T(\widehat{m}(\theta(\mu))) \ge Z_\alpha \right) f_\mu(\mu) d\mu = 1 - \alpha + o_p(1). \tag{17}$$

Similar to the above local case, for a given μ, under the posterior the distribution of $\widehat{m}(\theta) - \mu + \mu$ coincides with the asymptotic distribution of $\widehat{m}(\theta(\mu)) - \mu + \mu$ under $\mathbb{P}_\mu$, and thus Equation (17) is implied by (16).

□

7.3.1. *Step of Gaussian Integral.* In this section, we derive a lemma regarding the Gaussian integral and the tail probability involved in the above main theorem.

Lemma 2. Under Assumptions 7-8, $0 \le \kappa < \infty$, we have, for any fixed $\mu \in \Gamma$, $\varepsilon \asymp \sqrt{k} \log T$,

$$\int_{B_\varepsilon} N_T(\theta|\mu) \left| \left(1 + \|\theta - \theta(\mu)\|^\kappa \right) \right| \mathrm{d}\theta N_T(\mu) d\mu \lesssim_p k^{\kappa/2},$$

and

$$\sup_{\mu \in \Gamma} \mathbb{P}_\gamma (\| C_{w,\mu}^{-1/2} (\sqrt{T}^{-1} \gamma - C_{w,\mu}^{-1/2} C_{m,\mu}) \|^\kappa > \frac{\varepsilon}{\sqrt{T}} |\mu) \lesssim_p \sqrt{k}^\kappa \exp(-\varepsilon).$$

Proof. It suffices to just show it for $\kappa > 0$. It is well known that for a positive continuous random variable X, $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x) dx$. Recall that Γ_ϵ is the set of μ corresponds B_ε, where ε is set to be $\varepsilon \asymp \sqrt{k} \log T$. We also have the fact that $\int_{B_\varepsilon} d\theta d\mu \le \int_\Gamma \int_{B_{\varepsilon,\theta|\mu}} d\theta d\mu$, with

$B_{\varepsilon,\theta|\mu} \overset{\text{def}}{=} \{\theta : \sqrt{T}\|G(\theta(\nu(\mu)))(\theta-\theta(\nu(\mu)))\| \le \varepsilon\}$ for a point $\mu \in \Gamma_\epsilon$, therefore, we may start by looking at $\int_{B_{\varepsilon,\theta|\mu}} N_T(\theta|\mu)\left|\left(1+\|\theta-\theta(\nu(\mu))\|^\kappa\right)\right| \mathrm{d}\theta$. Since for $x \ge 0$, $x = \int_{z\ge 0} \mathbf{1}(z \le x)dz$,

$$\begin{aligned}
&\int_{B_{\varepsilon,\theta|\mu}} N_T(\theta|\mu)(1+\|\theta-\theta(\nu(\mu))\|^\kappa)\mathbf{1}(\theta\in B_{\varepsilon,\theta})d\theta \\
&= \int_{B_{\varepsilon,\theta|\mu}}\int_0^\infty \mathbf{1}(x \le 1+\|\theta-\theta(\nu(\mu))\|^\kappa)\mathbf{1}(\theta\in B_{\varepsilon,\theta})dx N_T(\theta|\mu)d\theta \\
&= \int_0^\infty\int_{B_{\varepsilon,\theta|\mu}} \mathbf{1}(1+\|\theta-\theta(\nu(\mu))\|^\kappa \ge x)\mathbf{1}(\theta\in B_{\varepsilon,\theta})N_T(\theta|\mu)d\theta dx.
\end{aligned}$$

As in the proof of Theorem 2, $\mathbb{P}_{N_T(\theta|\mu)}(.)$ is the probability measure conditioning on μ corresponding to the density $N_T(\theta|\mu)$, and γ is a standard k-dimensional multivariate Gaussian random variable with the associated probability measure $\mathbb{P}_\gamma$ and density $f(\lambda)$ with respect to the Lebesgue measure. Let us look for a fixed μ, then we can proceed from the above as follows,

$$\begin{aligned}
&\int_0^\infty \mathbb{P}_{N_T(\theta|\mu)}(\|\theta-\theta(\nu(\mu))\|^\kappa > x-1, \|\sqrt{T}G(\theta(\nu(\mu)))(\theta-\theta(\nu(\mu)))\|^2 \le \varepsilon^2|\mu)dx \\
&= \int_0^\infty \mathbb{P}_\gamma(\|\sqrt{T}^{-1}C_{w,\mu}^{-1/2}(\gamma - C_{w,\mu}^{-1/2}C_{m,\mu})\| > ((x-1)\vee 0)^{1/\kappa}, \\
&\quad \|\sqrt{T}^{-1}G(\theta(\nu(\mu)))(C_{w,\mu}^{-1/2}(\gamma - C_{w,\mu}^{-1/2}C_{m,\mu}))\|^2 \le \varepsilon^2/T|\mu)dx \\
&\le \int_0^\infty \mathbb{P}_\gamma(\|\sqrt{T}^{-1}C_{w,\mu}^{-1/2}\gamma\| > ((x-1)\vee 0)^{1/\kappa} - \|C_{w,\mu}^{-1}C_{m,\mu}\||\mu)dx.
\end{aligned}$$

Let $\lambda = \left(\sqrt{T}^{-1}\lambda_{\max}\left(C_{w,\mu}^{-1/2}\right)\right)^2$, and by Assumption 6, $\lambda \lesssim_p T^{-1}$, $\|C_{w,\mu}^{-1}C_{m,\mu}\| \lesssim_p \frac{\sqrt{k}}{\sqrt{T}}$. Note that,

$$\begin{aligned}
&\int_0^\infty \mathbb{P}_\gamma(\|\sqrt{T}^{-1}C_{w,\mu}^{-1/2}\gamma\| > ((x-1)\vee 0)^{1/\kappa} - \|C_{w,\mu}^{-1/2}C_{m,\mu}\||\mu)dx \\
&= \int_0^{1+(\sqrt{k}+\sqrt{T}\|C_{w,\mu}^{-1}C_{m,\mu}\|)^\kappa} \mathbb{P}_\gamma(\|\sqrt{T}^{-1}C_{w,\mu}^{-1/2}\gamma\| > ((x-1)\vee 0)^{1/\kappa} - \|C_{w,\mu}^{-1}C_{m,\mu}\||\mu)dx \\
&\quad + \int_{1+(\sqrt{k}+\sqrt{T}\|C_{w,\mu}^{-1}C_{m,\mu}\|)^\kappa}^\infty \mathbb{P}_\gamma(\|\sqrt{T}^{-1}C_{w,\mu}^{-1/2}\gamma\| > (x-1)^{1/\kappa} - \|C_{w,\mu}^{-1}C_{m,\mu}\||\mu)dx,
\end{aligned}$$

where the first term can be bounded by $\int_0^{1+(\sqrt{k}+\sqrt{T}\|C_{w,\mu}^{-1}C_{m,\mu}\|)^\kappa} 1dx \lesssim k^{\frac{\kappa}{2}}$, and the second term, via the inequality as in Lemma 7. For a positive constant $c > 0$, it can be bounded by $\int_{c(\sqrt{k})^\kappa}^\infty \exp(-z^{2/\kappa}/2)dz \lesssim_p k^{\frac{\kappa}{2}}$. The second statement in Lemma 2 is a direct result of Lemma 7. □

7.3.2. *Step of c^*.* In this subsection, we study the term of c^*,

$$c^* = \int_\Xi \exp(-\frac{1}{2}V_T(h(.),\theta,\mu) + \log\pi(\mu))d\theta d\mu / \int_\Xi \exp(\frac{1}{2}Q(\theta,\mu) + \log\pi(\theta,\mu))d\theta d\mu$$

Lemma 3. Under Assumptions 7-8, we have $1 \lesssim_p c^* \lesssim_p 1$.

Proof. Define
$$c_1^* = \int_{B_\varepsilon} \exp(-V_T(h(.),\theta,\mu)/2 + \log(\pi(\mu)))d\theta d\mu,$$
and
$$c_2^* = \int_{B_\varepsilon} \exp(\frac{1}{2}Q_T(\theta,\mu) + \log\pi(\mu,\theta))d\theta d\mu.$$
$$\begin{aligned}
c^* &= \frac{\int_\Xi \exp(-V_T(h(.),\theta,\mu)/2 + \log(\pi(\mu)))d\theta d\mu}{\int_\Xi \exp(\frac{1}{2}Q_T(\theta,\mu) + \log\pi(\mu,\theta))d\theta d\mu} \\
&= \frac{\int_{B_\varepsilon} \exp(-V_T(h(.),\theta,\mu)/2 + \log(\pi(\mu)))d\theta d\mu}{\int_{B_\varepsilon} \exp(\frac{1}{2}Q_T(\theta,\mu) + \log\pi(\mu,\theta))d\theta d\mu} + o_p(1) = \frac{c_1^*}{c_2^*} + o_p(1) \\
&= \frac{c_1^*}{\int_{B_\varepsilon} \exp(-V_T(h(.),\theta,\mu)/2 + \log(\pi(\mu)) + R_T(\theta,\mu))d\theta d\mu} + o_p(1) \\
&= \frac{c_1^*}{\int_{B_\varepsilon} \exp(-V_T(h(.),\theta,\mu)/2 + \log(\pi(\mu)))[\exp(R_T(\theta,\mu)) - 1]d\theta d\mu + c_1^*} \\
&+ o_p(1).
\end{aligned}$$

If the above term is of order $\frac{\int_{B_\varepsilon} \exp(-V_T(h(.),\theta,\mu)/2+\log(\pi(\mu)))d\theta d\mu}{\int_{B_\varepsilon} \exp(-V_T(h(.),\theta,\mu)/2+\log(\pi(\mu)))d\theta d\mu(1+o_p(1))} + o_p(1)$, then we reach the conclusion.

It boils down to show that $\sup_{\theta,\mu\in B_\varepsilon} |R_T(\theta,\mu)| = o_p(1)$. Because Assumption 8,
$$\sup_{(\theta,\mu)\in B_\varepsilon} T|R_T(\theta,\mu)|/(\|\sqrt{T}h(\theta,\mu)\|^2 + k(\log T)^2) \lesssim_p \frac{\sqrt{k}(\log T)^2}{\sqrt{T}} \vee [\frac{q}{\sqrt{kT}}] \to 0.$$
Then we have because of Assumption 8,
$$\sup_{(\theta,\mu)\in B_\varepsilon} |R_T(\theta,\mu)| \lesssim_p \frac{(k^{3/2}(\log T)^4)}{\sqrt{T}} \vee \frac{k(\log T)^2 q}{\sqrt{kT}} \to 0.$$
Finally we show that c_1^* has a rate.

□

7.3.3. *Steps of $R_T(\theta,\mu)$.* Define $M_{m,T}(\theta,\mu) = \sqrt{T}[\widehat{m}(\theta) - \widehat{m}(\theta(\nu(\mu)) - \mathbb{E}(\widehat{m}(\theta)) + \mathbb{E}(\widehat{m}(\theta(\nu(\mu)))]$.

Assumption 10. *(Tail assumptions of empirical moments) $\widehat{m}(\theta)$ and $\mathbb{E}(\widehat{m}(\theta))$ is second order differentiable in θ. Let $\gamma_1 \in \mathbb{R}^k$ and $\gamma_2 \in \mathbb{R}^q$, and γ_1 and γ_2 are unit vectors under $\|\cdot\|$, there exists constants $u, v_0 > 0$ such that $\sup_{\gamma_1,\gamma_2,(\theta,\mu)\in B_\varepsilon} \log\mathbb{E}\exp(\lambda\gamma_1^\top(\partial M_{m,T}(\theta,\mu)/\partial\theta)^\top\gamma_2) \lesssim v_0^2\lambda^2/2$, where $v_0 \lesssim \frac{1}{\sqrt{T}}$, $|\lambda| \le u$, and $u \ge \sqrt{k}$. Additionally, $\sup_{\mu\in\Gamma} |\widehat{m}(\theta(\mu)) - \mu|_{\max} \lesssim_p \frac{(\sqrt{k}\log T)}{\sqrt{T}}$.*

In Assumption 10, u represents the strength of the tail assumption. Furthermore, v_0 corresponds to a proxy for variance, while the assumption
$$\sup_{\gamma_1,\gamma_2,(\theta,\mu)\in B_\varepsilon} \log\mathbb{E}\exp(\lambda\gamma_1^\top(\partial M_{m,T}(\theta,\mu)/\partial\theta)^\top\gamma_2) \lesssim v_0^2\lambda^2/2$$

resembles a sub-Gaussian restriction, which remains valid under mild conditions. For instance, consider $G_T(\theta) = \frac{\partial \widehat{m}(\theta)}{\partial \theta}$. Consequently, based on the differentiable condition in Assumption 10, we have $\frac{\partial M_{m,T}(\theta,\mu)}{\partial \theta} = \frac{\partial[\widehat{m}(\theta) - \widehat{m}(\theta(\nu(\mu))) - \mathbb{E}(\widehat{m}(\theta)) + \mathbb{E}(\widehat{m}(\theta(\nu(\mu))))]}{\partial \theta} = G_T(\theta) - G(\theta))$. $G_T(\theta) - G(\theta)$ is a centered process, and it is reasonable to assume that it has sub-Gaussian tails.

Assumption 11. *(Moments assumptions and identification) $\mathbb{E}\widehat{m}(\theta)$ is second order continuously differentiable, and $\sup_{\theta\in\Theta} \lambda_{\max}\big(\partial^2_\theta \mathbb{E}[g(Z_t,\theta)]\big) \le C$. All singular values of $G\big(\theta\big)$ are bounded from above and away from zero for all $\theta \in \Theta$. Furthermore, for all $\theta \in \Theta$ $\|W(\theta) - W_T(\theta)\| = O_p\Big((\log T)^2\sqrt{\frac{k}{T}}\Big)$. for any $\varepsilon > 0$, $\sup_{\|\theta'-\theta\|\le\varepsilon} \|W_T(\theta') - W_T(\theta)\| = O_p\big(\varepsilon\big)$, and $\frac{k^2(\log T)^4}{T} \to 0$, $\frac{q^2}{kT} \to 0$.*

Lemma 4. Assumption 11 implies that $\mathrm{tr}((G(\theta)^\top G(\theta)^{-1/2}) \asymp k$. , $\lambda_{\max}(W(\theta))$ is bounded and there exists a positive constant c such that $\|c[G(\theta)^\top G(\theta)]\| \le 1$. Under Assumptions 7, 10 -11, we have the following:

$$\sup_{(\theta,\mu)\in B_\varepsilon} \Big|\frac{TR_T(\theta,\mu)|}{[k(\log T)^2 + \|\sqrt{T}h(\theta,\mu)\|^2]}\Big| \lesssim_p \frac{\sqrt{k}(\log T)^2}{\sqrt{T}} \vee \frac{q}{\sqrt{kT}}.$$

Proof. In this step, we verify the detailed derivation regarding $R_T(\theta,\mu)$ in relation to Assumption 8. The main goal of this derivation is to show that on B_ε, exists an arbitrary constant $\delta > 0$, such that with probability approaching 1,

$$\sup_{(\theta,\mu)\in B_\varepsilon} |TR_T(\theta,\mu)| \le \frac{\delta(\|\sqrt{T}h(\theta,\mu)\|^2 + k(\log T)^2)\sqrt{k}(\log T)^3}{\sqrt{T}}. \tag{18}$$

Denote $r_T(\theta,\mu) = \widehat{m}(\theta) - \mu - (\widehat{m}(\theta(\nu(\mu))) - \mu) - h(\theta,\mu)$. We see that,

$$\begin{aligned}
&-T\left(\widehat{m}(\theta)-\mu\right)^\top W(\theta(\mu))\left(\widehat{m}(\theta)-\mu\right)\\
=&-T\left((\widehat{m}(\theta(\nu(\mu)))-\mu)+h(\theta,\mu)+r_T(\theta,\mu)\right)^\top W(\theta(\mu))\left((\widehat{m}(\theta(\nu(\mu)))-\mu)+h(\theta,\mu)+r_T(\theta,\mu)\right),
\end{aligned}$$

and thus

$$\begin{aligned}
TR_T(\theta,\mu) =& -T\left(r_T(\theta,\mu)\right)^\top W(\theta(\mu))\left(2h(\theta,\mu)+r_T(\theta,\mu)\right) - 2T\left(r_T(\theta,\mu)\right)^\top W(\theta(\mu))\left((\widehat{m}(\theta(\nu(\mu)))-\mu)\right)\\
&+\log(\pi(\theta|\mu)) - \log(\pi(\theta(\nu(\mu))|\mu)) + T\left(\widehat{m}(\theta)-\mu\right)^\top (W(\theta(\mu)) - \widehat{W}_T^{-1}(\theta))\left(\widehat{m}(\theta)-\mu\right).
\end{aligned}$$

We see that $\|(\widehat{m}(\theta)-\mu)\| \lesssim \|r_T(\theta,\mu)\| + \|\widehat{m}(\theta(\mu)) - \mu\| + \|h(\theta,\mu)\| \lesssim_p \left(\frac{\sqrt{q}}{\sqrt{T}} \vee \frac{\sqrt{k}\log T\log T}{\sqrt{T}}\right)$. Thus, we have $\sup_{(\theta,\mu)\in B_\varepsilon} T\left(\widehat{m}(\theta)-\mu\right)^\top (W(\theta(\mu)) - \widehat{W}_T(\theta))\left(\widehat{m}(\theta)-\mu\right) \lesssim_p \left(\frac{q}{\sqrt{kT}} \vee \frac{(\log(T))^2\sqrt{k}}{\sqrt{T}}\right)\sup_{(\theta,\mu)\in B_\varepsilon}(\sqrt{k} + \sqrt{T}\|h(\theta,\mu)\|)$.

By Assumption 11, we have that $W(\theta(\mu))$ has a bounded maximum eigenvalue, and from Lemma 5 $\sup_{(\theta,\mu)\in B_\varepsilon} r_T(\theta,\mu) \lesssim_p \sup_{(\theta,\mu)\in B_\varepsilon} \frac{\sqrt{k}\left(\sqrt{k}+\sqrt{T}\|h(\theta,\mu)\|\right)}{T}$, therefore, implied by Lemma 5 we have the following on B_ε,

$$
\begin{aligned}
&\sup_{(\theta,\mu)\in B_\varepsilon} TR_T(\theta,\mu)\\
\lesssim_p &\sup_{(\theta,\mu)\in B_\varepsilon} \lambda_{\max}(W(\theta(\mu)))T[\|r_T(\theta,\mu)\|^2 \vee (\|r_T(\theta,\mu)\|\|h(\theta,\mu)\|)]\\
&+\sup_{(\theta,\mu)\in B_\varepsilon} \lambda_{\max}(W(\theta(\mu)))T[\|r_T(\theta,\mu)\|\|\widehat{m}(\theta(\nu(\mu)))-\mu)\|\\
&+\sup_{(\theta,\mu)\in B_\varepsilon} T\left(\widehat{m}(\theta)-\mu\right)^\top\left(W(\theta(\mu))-\widehat{W}_T^{-1}(\theta)\right)\left(\widehat{m}(\theta)-\mu\right)\\
\lesssim_p &\sup_{(\theta,\mu)\in B_\varepsilon} T(\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|)^2\left(\frac{\sqrt{k}}{T}\right)^2 \vee \sup_{(\theta,\mu)\in B_\varepsilon}\left(\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|\right)\frac{\sqrt{k}\varepsilon}{\sqrt{T}}\\
&+\sup_{(\theta,\mu)\in B_\varepsilon} \frac{T\sqrt{k}(\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|)}{T}\left(\frac{\sqrt{q}}{\sqrt{T}}\vee\frac{\sqrt{k}\log T}{\sqrt{T}}\right)\\
&+\frac{T(\log T)^2\sqrt{k}}{\sqrt{T}}\left(\frac{\sqrt{q}}{\sqrt{T}}\vee\frac{\sqrt{k}\log T}{\sqrt{T}}\right)^2.
\end{aligned}
$$

Thus,

$$\sup_{\theta,\mu\in B_\varepsilon}\frac{\left|TR_T(\theta,\mu)\right|}{k(\log T)^2+\|\sqrt{T}h(\theta,\mu)\|^2}\lesssim_p\frac{(\log T)^2\sqrt{k}}{\sqrt{T}}.$$

Additionally, on B_ε for sufficiently large T, we have that

$$\sup_{(\theta,\mu)\in B_\varepsilon}|\exp(R_T(\theta,\mu))-1|\le\sup_{(\theta,\mu)\in B_\varepsilon}\left(|R_T(\theta,\mu)|+|R_T(\theta,\mu)|^2\right),$$

which is due to fact that $|e^x-1-x-\frac{1}{2}x^2|=o(|x|^2)$ for sufficiently small $|x|$.

□

7.3.4. *Proof of $r_T(\theta,\mu)$.* This section provides the intermediate results used in the proof of Lemma 4

Lemma 5. Under Assumptions 10 to11,

$$\sup_{\theta,\mu\in B_\varepsilon}\frac{\|r_T(\theta,\mu)\|}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}\lesssim_p\frac{\sqrt{k}}{T},$$

Proof. Let $\mathscr{M}_{m,T}(\theta,\mu)=\frac{M_{m,T}(\theta,\mu)}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}$, which is a centered object by definition and satisfies the following, $\mathbb{E}\left(\mathscr{M}_{m,T}(\theta,\mu)\right)=0$. We focus on its behaviors on B_ε. Assumption 11 imposes differentiability conditions on $\mathbb{E}(\widehat{m}(\theta))-\mathbb{E}(\widehat{m}(\theta(\nu(\mu)))$ concerning θ, and thus, the residual terms $\mathscr{M}_{m,T}(\theta,\mu)$ shrinks to 0 within B_ε because, on B_ε, the higher-order terms related to $\|G(\theta(\nu(\mu)))(\theta-\theta(\nu(\mu)))\|$ also diminishes to 0 following Assumption 11. We use $\mathscr{M}_{m,T}(\theta,\mu)$ to analyze $r_T(\theta,\mu)$.

By construction, we have that $\frac{r_T(\theta,\mu)}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}-\mathscr{M}_{m,T}(\theta,\mu)=\frac{-h(\theta,\mu)+\mathbb{E}(\widehat{m}(\theta))-\mathbb{E}(\widehat{m}(\theta(\nu(\mu))))}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}$. According to Assumption 11, on B_ε, $\|\frac{r_T(\theta,\mu)}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}-\mathscr{M}_{m,T}(\theta,\mu)\|$ is bounded by

$$\frac{\|-h(\theta,\mu)+(\sup_{\theta,\mu\in B_\varepsilon}\lambda_{\max}(\partial^2\mathbb{E}(g(Z_t,\theta))/\partial\theta\partial\theta^\top))\|\theta-\theta(\nu(\mu))\|^2\|}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}\lesssim\frac{\sqrt{k}[\sqrt{k}\log(T)]^2}{T}.$$

Assumption 11 implies that $\frac{k\log(T)}{\sqrt{T}}\to 0$, and thus $\sup_{\theta,\mu\in B_\varepsilon}\|\frac{r_T(\theta,\mu)}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}-\mathscr{M}_{m,T}(\theta,\mu)\|\ll\frac{\sqrt{k}(\log T)}{\sqrt{T}}$ on B_ε. Then it suffices to examine $\mathscr{M}_{m,T}(\theta,\mu)$ for the rate of $\frac{r_T(\theta,\mu)}{\sqrt{k}\log T+\sqrt{T}\|h(\theta,\mu)\|}$. We then use Lemma 8 to analyze the term $M_{m,T}(\theta,\mu)$, where the variable μ in Theorem B.15 corresponds to $(\theta-\theta(\nu(\mu)))\in\mathbb{R}^k$ in our context. As a result, $\Upsilon_\circ(r)$ should represent a ball containing B_ε.

We next analyze the terms, i.e., A and $v_0 r\mathfrak{z}_{\mathbb{H}}(\mathrm{x})$ in Theorem B.15 in Spokoiny (2017), where v_0 refers to variance, r refers to radius and $\mathfrak{z}_{\mathbb{H}}(\mathrm{x})$ refers to entropy of $\Upsilon_\circ(r)$. By Assumption 11, we have $r\approx\frac{\sqrt{k}\log T}{\sqrt{T}}$. The A therein is identity I_q, and $\mathfrak{z}_{\mathbb{H}}(\mathrm{x})\approx\sqrt{k}$, and due to the second differentiability assumption in Assumption 10, $v_0\lesssim 1/\sqrt{T}$.

Then, we have $\sup_{\theta,\mu\in B_\varepsilon}M_{m,T}(\theta,\mu)\lesssim_P\frac{k\log T}{T}$ and $\sup_{\theta,\mu\in B_\varepsilon}\mathscr{M}_{m,T}(\theta,\mu)\lesssim_P\frac{\sqrt{k}}{T}$. □

7.3.5. *Useful Lemmas.* Here, we list a few useful lemmas from Spokoiny (2017) and Spokoiny and Panov (2019).

Lemma 6. (Corollary A.3. from Spokoiny and Panov (2019)) Let γ be a standard normal random vector in $\mathbb{R}^k$. Then for any $\mathrm{x}>0$

$$\mathbb{P}\left(\|\gamma\|^2\geq k+2\sqrt{k\mathrm{x}}+2\mathrm{x}\right)\leq \mathrm{e}^{-\mathrm{x}},$$
$$\mathbb{P}(\|\gamma\|\geq\sqrt{k}+\sqrt{2\mathrm{x}})\leq\mathrm{e}^{-\mathrm{x}},$$
$$\mathbb{P}\left(\|\gamma\|^2\leq k-2\sqrt{k\mathrm{x}}\right)\leq\mathrm{e}^{-\mathrm{x}}.$$

Lemma 7. (Theorem A.2. from Spokoiny and Panov (2019)) Let H be a positive definite matrix. Let $\boldsymbol{\xi}\sim N\left(0,H^2\right)$ be a mean-zero normal random vector in $\mathbb{R}^k$ and B be a symmetric non-negative definite matrix such that $A=H^{-1}BH$ is a trace operator in $\mathbb{R}^k$. Then with $k=\operatorname{tr}(A)$, $\mathrm{v}^2=\operatorname{tr}\left(A^2\right)$, and $\lambda=\|A\|$, it holds for each $\mathrm{x}\geq 0$,

$$\mathbb{P}\left(\boldsymbol{\xi}^\top B\boldsymbol{\xi}\geq z^2(A,\mathrm{x})\right)\leq\mathrm{e}^{-\mathrm{x}},$$
$$\text{with}\quad z(A,\mathrm{x})\stackrel{\text{def}}{=}\sqrt{\mathrm{k}+2\mathrm{v}\mathrm{x}^{1/2}+2\lambda\mathrm{x}}.$$

It also implies

$$\mathbb{P}\left(\left\|B^{1/2}\boldsymbol{\xi}\right\|>\mathrm{k}^{1/2}+(2\lambda\mathrm{x})^{1/2}\right)\leq\mathrm{e}^{-\mathrm{x}}.$$

If B is symmetric but not necessarily positive, then

$$\mathbb{P}\left(|\boldsymbol{\xi}^\top B\boldsymbol{\xi}-\mathrm{k}|>2\mathrm{v}\mathrm{x}^{1/2}+2\lambda\mathrm{x}\right)\leq 2\mathrm{e}^{-\mathrm{x}}.$$

Lemma 8. (Theorem B.15. from Spokoiny (2017)) Let $\mathscr{Y}(\boldsymbol{v})$ with $\boldsymbol{v} \in \Upsilon_0(r) = \{\boldsymbol{v} \in \Upsilon : \|\boldsymbol{v} - \boldsymbol{v}^*\| \le r\}$ and $\Upsilon \subseteq \mathbb{R}^k$, be a smooth centered random vector process with values in $\mathbb{R}^q$. Let also $\mathbb{E}[\mathscr{Y}(\boldsymbol{v}^*)] = 0$ for the center $\boldsymbol{v}^* \in \Upsilon_0(r)$. Without loss of generality, assume $\boldsymbol{v}^* = 0$. We aim to bound $\|\mathscr{Y}(\boldsymbol{v})\|$ uniformly over $\boldsymbol{v}$ over a vicinity $\Upsilon_0(r)$ of $\boldsymbol{v}^*$. By $\nabla\mathscr{Y}(\boldsymbol{v})$ we denote the $k \times q$ matrix with entries $\nabla_{\boldsymbol{v}_i}\mathscr{Y}(\boldsymbol{v}), i \le k, j \le q$. Suppose that $\mathscr{Y}(\boldsymbol{v})$ satisfies for each $\boldsymbol{\gamma}_1 \in \mathbb{R}^k$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^q$ with $\left\|\boldsymbol{\gamma}_1\right\| = \left\|\boldsymbol{\gamma}_2\right\| = 1$, and there exists a positive constant v_0,

$$\sup_{\boldsymbol{v}\in\Upsilon} \log \mathbb{E} \exp\left\{\lambda \boldsymbol{\gamma}_1^\top \nabla\mathscr{Y}(\boldsymbol{v})\boldsymbol{\gamma}_2\right\} \le \frac{v_0^2\lambda^2}{2}, \quad |\lambda| \le g.$$

Let A be a matrix fulfilling $1/2 \le \left\|AA^\top\right\| \le 1$. Then for each r, it holds

$$\mathbb{P}\left\{\sup_{\boldsymbol{v}\in\Upsilon_0(r)} \|A\nabla\mathscr{Y}(\boldsymbol{v})\| > \sqrt{8}v_0 r \mathfrak{z}_{\mathbb{H}}(\mathrm{x})\right\} \le \mathrm{e}^{-\mathrm{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathrm{x})$ is given by the following with $\mathbb{Q}_2 = p_A + \mathbb{Q}_2(\Upsilon_\circ(r))$.

$$\mathfrak{z}_{\mathbb{H}}(\mathrm{x}) = \begin{cases} 2\sqrt{\mathbb{Q}_2 + 2\mathrm{x}}, & \text{if } \mathbb{Q}_2 + 2\mathrm{x} \le \mathrm{g}^2, \\ 2\,\mathrm{g}^{-1}\mathrm{x} + \mathrm{g}^{-1}\mathbb{Q}_2 + \mathrm{g}, & \text{if } \mathbb{Q}_2 + 2\mathrm{x} > \mathrm{g}^2. \end{cases}$$

In the above, $\mathbb{Q}_2$ relates to the entropy of the set $\Upsilon_0(r)$, p_A denotes a trace norm of a trace operator A^{-2}, and both can be calculated according to section B.4 in Spokoiny (2017) as outlined below.

For each $k \le 1$, by $\mathscr{M}_k$ we denote a r_k-net in $\Upsilon^\circ(r_0)$ with $r_k = r_0 2^{-k}$, so that $\Upsilon^\circ(r_0) \subseteq \bigcup_{\boldsymbol{v}\in\mathscr{M}_k}\{\boldsymbol{v}' \in \Upsilon : \|\boldsymbol{v}' - \boldsymbol{v}\| \le r_k\}$, then $\mathbb{Q}_2(\Upsilon^\circ) \stackrel{\text{def}}{=} \sum_{k=1}^\infty 2^{-k+1}\log(2\mathbb{N}_k)$ with $\mathbb{N}_k \stackrel{\text{def}}{=} |\mathscr{M}_k|$ being the cardinality of $\mathscr{M}_k$. For a positive self-adjoint operator in $\mathbb{R}^\infty$, denoted by $\mathbb{H}$, such that $\lambda_{\min}(\mathbb{H}) = 1$ and $\mathbb{H}^{-2}$ is a trace operator, then $\mathrm{p}_{\mathbb{H}} \stackrel{\text{def}}{=} \operatorname{tr}\left(\mathbb{H}^{-2}\right) = \sum_{j=1}^\infty h_j^{-2} < \infty$, where $1 = h_1 \le h_2 \le \cdots$ are the ordered eigenvalues of H.

Lemma 9. (Theorem B.8. in Spokoiny (2017)) Suppose that for some $\alpha > 1$, $\mathrm{p}_{\mathbb{H}}(\alpha) \stackrel{\text{def}}{=} \sum_{j=1}^\infty h_j^{-2}\log^\alpha\left(h_j^2\right) < \infty$, then $\mathbb{Q}_2(\Upsilon_{\mathbb{H}}^\circ(r)) \le \mathrm{C}\sum_{j=1}^\infty h_j^{-1}$, with $\Upsilon_{\mathbb{H}}^\circ(r) = \{\boldsymbol{v}' \in \Upsilon : \|\mathbb{H}(\boldsymbol{v}' - \boldsymbol{v})\| \le r\}$, for a fixed center $\boldsymbol{v}$.

Lemma 10. (Lemma A.17 from Spokoiny and Panov (2019)) Let $\mathscr{T}$ be a linear operator in $\mathbb{R}^k$ with $\|\mathscr{T}\|_{\mathrm{op}} \le 1$. Let $\boldsymbol{z} \in \mathbb{R}^k$ be a unit norm vector: $\|\boldsymbol{z}\| = 1$. Define $k = \operatorname{tr}\left(\mathscr{T}^\top\mathscr{T}\right)$. For any positive $\mathrm{C}_0, \mathrm{r}_0$ with $1/2 < \mathrm{C}_0 \le 1$ and $\mathrm{C}_0\mathrm{r}_0 > 2\sqrt{k+1} + \sqrt{\mathrm{x}}$,

$$\mathbb{E}\left\{|\langle\boldsymbol{z},\gamma\rangle|^2 \exp\left(-\mathrm{C}_0\mathrm{r}_0\|\mathscr{T}\gamma\| + \frac{\mathrm{C}_0\mathrm{r}_0^2}{2} + \frac{1}{2}\|\mathscr{T}\gamma\|^2\right)\mathbb{I}\left(\|\mathscr{T}\gamma\| > \mathrm{r}_0\right)\right\} \le \mathrm{C}\mathrm{e}^{-(k+\mathrm{x})/2}.$$

7.4. **Proof of Theorem 3.**

Proof.

Step 1

Under the information equality, we have $\sigma^2_{\eta,\mu} = (\partial\eta(\theta(\mu))/\partial\theta)^\top J_W(\theta(\mu))^{-1}(\partial\eta(\theta(\mu))/\partial\theta)$. From Assumption 9, the frequentist confidence interval is presented as

$$[(\partial\eta(\theta(\mu))/\partial\theta)^\top\widehat{\theta}(\mu) + \frac{\sigma_{\eta,\mu}z_{\alpha/2}}{\sqrt{T}}, (\partial\eta(\theta(\mu))/\partial\theta)^\top\widehat{\theta}(\mu) + \frac{z_{1-\alpha/2}\sigma_{\eta,\mu}}{\sqrt{T}}].$$

Step 2

As the second step, we need to prove that under the information equality, the confidence interval agrees with the frequentist confidence interval. As suggested by Theorem 2, $p_T(\theta,\mu)$ can be approximated well by $N_T(\theta,\mu)$, and the conditional distribution of θ on μ (with density $N_T(\theta|\mu)$) follows a Gaussian distribution with mean $C^{-1}_{w,\mu}C_{m,\mu}$ and variance $(TC_{w,\mu})^{-1}$. As we noticed, $\partial\eta(\theta(\mu))/\partial\theta^\top(TC_{w,\mu})^{-1}\partial\eta(\theta(\mu))/\partial\theta = \sigma^2_{\eta,\mu}/T$ under information equality. Thus, it suffices to check that the quantiles of $p_T(\theta,\mu)$ and $N_T(\theta,\mu)$ indeed agree.

Let

$$\begin{aligned}
H_{\eta,T}(s,\mu) =& F_{\eta,T}\left(\eta\left(\theta(\mu)\right) + s/\sqrt{T}\right) = \int_{\theta\in\Theta:\eta(\theta)\le\eta\left(\theta(\mu)\right)+s/\sqrt{T}} p_T(\theta,\mu)/p_T(\mu)d\theta,\\
\widehat{H}_{\eta,T}(s,\mu) =& \int_{\theta\in\Theta:\eta(\theta)\le\eta\left(\theta(\mu)\right)+s/\sqrt{T}} N_T(\theta,\mu)/N_T(\mu)d\theta,\\
H_{\eta,\infty}(s,\mu) =& \int_{\theta\in\Theta:(\partial\eta(\theta(\mu))/\partial\theta)^\top(\theta-\theta(\mu))\le s/\sqrt{T}} N_T(\theta,\mu)/N_T(\mu)d\theta,
\end{aligned}$$

where $N_T(\mu) = \int_{\theta\in\Theta} N_T(\theta,\mu)d\theta$. By definition of total variation of moments norm and Theorem 2, we have

$$\sup_{s\in\mathscr{S}(\mu),\mu\in\Gamma}\left|H_{\eta,T}(s,\mu) - \widehat{H}_{\eta,T}(s,\mu)\right| \to_p 0,$$

where $\mathscr{S}(\mu)$ denotes the support of $H_{\eta,T}(\cdot,\mu)$ such that $\mathscr{S}(\mu) = \{s\in\mathbb{R} : s = \sqrt{T}(\eta(\theta)-\eta(\theta(\mu))), \theta\in\Theta\}$.

By the uniform continuity of the integral of the normal density with respect to the boundary integration, we have

$$\sup_{s\in\mathscr{S}(\mu),\mu\in\Gamma}\left|\widehat{H}_{\eta,T}(s,\mu) - H_{\eta,\infty}(s,\mu)\right| \to_p 0,$$

which implies that

$$\sup_{s\in\mathscr{S}(\mu),\mu\in\Gamma}\left|H_{\eta,T}(s,\mu) - H_{\eta,\infty}(s,\mu)\right| \to_p 0.$$

The convergence of the distribution function implies the convergence of quantiles at continuous points of distribution functions so that $H^{-1}_{\eta,T}(\alpha,\mu) - H^{-1}_{\eta,\infty}(\alpha,\mu) \to_p 0$, where $H^{-1}_{\eta,\infty}(\alpha,\mu)$ and $H^{-1}_{\eta,\infty}(\alpha,\mu)$ are defined as the inverse of the function $H_{\eta,T}(s,\mu)$ in terms of s for any fixed μ.

Next, similar to the proof of Theorem 3 in Chernozhukov and Hong (2003) and we have that

$$H_{\eta,\infty}(s,\mu) = \mathbb{P}_{N_T(\theta|\mu)}\left\{(\partial\eta(\theta(\mu))/\partial\theta)^\top(\theta-\theta(\mu)) \le s/\sqrt{T}\right\}$$

so that $H^{-1}_{\eta,\infty}(\alpha,\mu) = (\partial\eta(\theta(\mu))/\partial\theta)^\top\sqrt{T}\tilde{U}_T(\mu) + q_\alpha\sqrt{(\partial\eta(\theta(\mu))/\partial\theta)^\top J_W(\mu)^{-1}(\partial\eta(\theta(\mu))/\partial\theta)}$ implied by the proof of Theorem 2, where q_α is the α-quantile of a standard normal distribution.

The rest of the results follow from the fact that $H_{\eta,T}^{-1}(\alpha,\mu)=\sqrt{T}(c_{\eta,T}(\alpha,\mu)-\eta(\theta(\mu)))$ and the delta method.

Recall that $f_\mu(\mu)=\partial F_\mu(\mu)/\partial\mu$ be the density corresponding to $\mathbb{P}^*(.)$. To prove the second statement, we have,

$$
\begin{aligned}
&\lim_{T\to\infty}\mathbb{P}^*\left\{\eta(\theta(\tilde{\mu}))\in\cup_{\mu'\in\mathcal{M}}\tilde{In}_{\alpha,\eta(.)}(\mu'),\forall\tilde{\mu}\in\Gamma\right\}\\
&=\lim_{T\to\infty}\int_\mu\mathbb{P}^*_{\theta(\mu),\mu}\left\{\eta(\theta(\tilde{\mu}))\in\cup_{\mu'\in\mathcal{M}}\tilde{In}_{\alpha,\eta(.)}(\mu'),\forall\tilde{\mu}\in\Gamma\right\}f_\mu(\mu)d\mu\\
&\geq\lim_{T\to\infty}\int_\mu\mathbb{P}^*_{\theta(\mu),\mu}\left\{\eta(\theta(\tilde{\mu}))\in\tilde{In}_{\alpha,\eta(.)}(\tilde{\mu}),\forall\tilde{\mu}\in\Gamma\right\}f_\mu(\mu)d\mu=1-\alpha.
\end{aligned}
$$

□

7.5. **Proof of Theorem 4.**

Proof. Keep $\mu\in\Gamma$ throughout this proof. In view of Assumption of the consistency of $\Omega_T(\theta(\mu))$, it suffices to show that $\|\widehat{J}_T^{-1}\left(\theta(\mu)\right)-J_W^{-1}(\theta(\mu))\|\to_p 0$, and then conclude using the delta method. Let $\zeta_\mu(\theta)$ be a function of θ such that

$$
\zeta_\mu(\theta)=\sqrt{T}\left(\theta-\theta(\mu)\right)-\sqrt{T}\underbrace{J_W\left(\theta(\mu)\right)^{-1}\Delta_{T,W}\left(\theta(\mu)\right)}_{\tilde{U}_T(\mu)},
$$

and the localized quasi-posterior density for $\zeta_\mu(\theta)$ is

$$
p_T(\zeta_\mu(\theta),\mu)=\frac{1}{\sqrt{T}}p_T\left(\zeta_\mu(\theta)/\sqrt{T}+\theta(\mu)+\tilde{U}_T(\mu),\mu\right).
$$

And similarly, define $N_T(\zeta_\mu(\theta),\mu)=\frac{1}{\sqrt{T}}N_T\left(\zeta_\mu(\theta)/\sqrt{T}+\theta(\mu)+\tilde{U}_T(\mu),\mu\right)$. Define H_T for the set of $\zeta_\mu(\theta)$ containing the set $\{\theta:\sqrt{T}\|h(\theta,\mu)\|\leq\varepsilon,\theta\in\Theta\}$. Denote $\zeta_\mu(\theta)=\left(\zeta_{\mu,1}(\theta),\ldots,\zeta_{\mu,k}(\theta)\right)$ and $\widetilde{T}_T=\left(\widetilde{T}_{T1},\ldots,\widetilde{T}_{Tk}\right)$ where $\widetilde{T}_T=\sqrt{T}\left(\widehat{\theta}(\mu)-\theta(\mu)\right)-\sqrt{T}\tilde{U}_T(\mu)$.

Note also

$$
\begin{aligned}
\widehat{J}_T^{-1}\left(\widehat{\theta}(\mu)\right)=&\int_\Theta T(\theta-\widehat{\theta}(\mu))(\theta-\widehat{\theta}(\mu))^\top p_T(\theta,\mu)/p_T(\mu)d\theta\mathbf{1}(\mu:p_T(\mu)>c)\\
=&\int_{H_T}\left(\zeta_\mu(\theta)-\sqrt{T}\left(\widehat{\theta}(\mu)-\theta(\mu)\right)+\sqrt{T}\tilde{U}_T(\mu)\right)\\
&\cdot\left(\zeta_\mu(\theta)-\sqrt{T}\left(\widehat{\theta}(\mu)-\theta(\mu)\right)+\sqrt{T}\tilde{U}_T(\mu)\right)^\top[p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)]d\zeta_\mu(\theta)\mathbf{1}(\mu:p_T(\mu)>c),
\end{aligned}
$$

and

$$
\tilde{J}_T^{-1}(\theta(\mu))\equiv\int_{H_T}\zeta_\mu(\theta)\zeta_\mu(\theta)^\top p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta).
$$

$$\tilde{J}_T^{-1,c}\left(\theta(\mu)\right) \equiv \int_{H_T} \zeta_\mu(\theta)\zeta_\mu(\theta)^\top p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta)\mathbf{1}(\mu : p_T(\mu) > c).$$

Therefore we have $\tilde{J}_T^{-1}(\theta(\mu)) - \tilde{J}_T^{-1,c}(\theta(\mu)) = \int_{H_T} \zeta_\mu(\theta)\zeta_\mu(\theta)^\top[p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)]d\zeta_\mu(\theta)\mathbf{1}(\mu : p_T(\mu) \le c)$. Due to the condition that $\int 1_{\{p_T(\mu)\le c\}} f_\mu(\mu)d\mu = o_p(1)$, thus we have $\|\tilde{J}_T^{-1}(\theta(\mu)) - \tilde{J}_T^{-1,c}(\theta(\mu))\| = o_p(1)$.

So $\widehat{J}_T^{-1}(\widehat{\theta}(\mu)) - \tilde{J}_T^{-1,c}(\theta(\mu)) = -2\int_{H_T} \zeta_\mu(\theta)\widetilde{T}_T^\top p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta)\mathbf{1}(\mu : p_T(\mu) > c) + \int_{H_T} \widetilde{T}_T\widetilde{T}_T^\top p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta)\mathbf{1}(\mu : p_T(\mu) > c)$, which will be verified to be ignorable in statement (c), (d), (e) and (f).

Denote $\Theta(\zeta_\mu(\theta))$ as the set corresponding to H_T. Recall that

$$J_W^{-1}\left(\theta(\mu)\right) = \int_{\Theta(\zeta_\mu(\theta))} \zeta_\mu(\theta)\zeta_\mu(\theta)^\top N_T(\zeta_\mu(\theta),\mu)/N_T(\mu)d\zeta_\mu(\theta) + o_p(1).$$

So

$$\begin{aligned} J_W^{-1}(\theta(\mu)) - \tilde{J}_T^{-1}(\theta(\mu)) = \quad & \int_{H_T} \zeta_\mu(\theta)\zeta_\mu(\theta)^\top (N_T(\zeta_\mu(\theta),\mu)/N_T(\mu) - p_T(\zeta_\mu(\theta),\mu)/p_T(\mu))d\zeta_\mu(\theta) \\ & + \int_{H_T^c} \zeta_\mu(\theta)\zeta_\mu(\theta)^\top p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta). \end{aligned}$$

This is verified in (a) and (b). Denote $\widetilde{T}_{Tj}$ as the jth element of $\widetilde{T}_T$ $(1 \le j \le k)$.

(a) $[\sum_{i,j}\{\int_{H_T} \zeta_{\mu,i}(\theta)\zeta_{\mu,j}(\theta)(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu) - N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))d\zeta_\mu(\theta)\}^2]^{1/2} = o_p(1)$ results from the following steps. First, note that one vector norm inequality that $|v|_2 = (\sum_i v_i^2)^{\frac{1}{2}} \le |v|_1 = (\sum_i |v_i|)$, $v \in \mathbb{R}^d$ implies the above term is upper bounded by

$$\sum_{i,j}\left|\int_{H_T} \zeta_{\mu,i}(\theta)\zeta_{\mu,j}(\theta)\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu) - N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right| d\zeta_\mu(\theta)\right|.$$

For each absolute term within the summation, we see by the Cauchy-Schwartz inequality or the Hölder inequality,

$$\begin{aligned} & \left|\int_{H_T} \zeta_{\mu,i}(\theta)\zeta_{\mu,j}(\theta)\left|p_T(\zeta_\mu(\theta),\mu)/p_T(\mu) - N_T(\zeta_\mu(\theta))\right| d\zeta_\mu(\theta)\right| \\ \le & \left[\int_{H_T} \zeta_{\mu,i}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu) - N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right| d\zeta_\mu(\theta)\right]^{1/2} \\ & \left[\int_{H_T} \zeta_{\mu,j}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu) - N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right| d\zeta_\mu(\theta)\right]^{1/2}. \end{aligned}$$

Then we have,

$$\sum_{i,j}\left[\int_{H_T}\zeta_{\mu,i}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right|d\zeta_\mu(\theta)\right]^{1/2}$$
$$\left[\int_{H_T}\zeta_{\mu,j}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right|d\zeta_\mu(\theta)\right]^{1/2}$$
$$=\sum_{i}\left[\int_{H_T}\zeta_{\mu,i}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right|d\zeta_\mu(\theta)\right]^{1/2}$$
$$\sum_{j}\left[\int_{H_T}\zeta_{\mu,j}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right|d\zeta_\mu(\theta)\right]^{1/2}.$$

We know

$$\sum_{i}\left[\int_{H_T}\zeta_{\mu,i}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right|d\zeta_\mu(\theta)\right]^{1/2}$$
$$\le k\left[\max_i\int_{H_T}\zeta_{\mu,i}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right|d\zeta_\mu(\theta)\right]^{1/2}$$

Then it boils down to bound $k\max_i\int_{H_T}\zeta_{\mu,i}(\theta)^2\left|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right|d\zeta_\mu(\theta)$, which is guaranteed by the rate assumption in Assumption 8.

(b) Next we look at $\sum_{1\le i,j\le k}|\int_{H_T^c}\zeta_{\mu,i}(\theta)\zeta_{\mu,j}(\theta)p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta)|=o_p(1)$ by definition of $p_T(\theta,\mu)$ and $J_T\left(\theta(\mu)\right)$ being uniformly nonsingular by Lemma 10. It suffice to look at $k|\int_{H_T^c}\sum_i h_i^2(\mu)p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta)|$. This is further upper bounded by

$$\left|\int_{H_T^c}\sum_i\zeta_{\mu,i}^2(\theta)p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)d\zeta_\mu(\theta)\right|$$
$$\le\left|\int_{H_T^c}\sum_i\zeta_{\mu,i}^2(\theta)|p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu)|d\zeta_\mu(\theta)\right|$$
$$+\left|\int_{H_T^c}\sum_i\zeta_{\mu,i}^2(\theta)N_T(\zeta_\mu(\theta),\mu)/N_T(\mu)]d\zeta_\mu(\theta)\right|$$

(c) By Assumption 9, we verify that $\|\widetilde{T}_T\|^2$ is of small order cite the rate of CUE. Thus, we have the $\int_{H_T}\underbrace{\|\widetilde{T}_T\|^2}_{=o_p((q/\sqrt{T})^2)}|\left((p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))\right)|d\zeta_\mu(\theta)=o_p((\frac{q}{\sqrt{T}})^2)$ by Theorem 2 ,

(d) $\int_{H_T}\underbrace{\|\widetilde{T}_T\|^2}_{=o_p((q/\sqrt{T})^2)}N_T(\zeta_\mu(\theta),\mu)/N_T(\mu)d\zeta_\mu(\theta)=o_p(1)$ by Theorem 2, definition of $N_T(\zeta_\mu(\theta),\mu)$, and $J_T\left(\theta(\mu)\right)$ being nonsingular,

(e) For all $1\le i,j\le k$, $\sum_{i,j}|\int_{H_T}\zeta_{\mu,i}(\theta)\underbrace{\widetilde{T}_{Tj}}_{=o_p(1/\sqrt{T})}|(p_T(\zeta_\mu(\theta),\mu)/p_T(\mu)-N_T(\zeta_\mu(\theta),\mu)/N_T(\mu))|d\zeta_\mu(\theta)|=$ $o_p(1)$ by Theorem 2 ,

(f) For all $1 \leq i, j \leq k$, $\sum_{i,j} | \int_{H_T} \zeta_{\mu,i}(\theta) \underbrace{\widetilde{T}_{Tj}}_{=o_p(1/\sqrt{T})} N_T(\zeta_\mu(\theta), \mu)/N_T(\mu) d\zeta_\mu(\theta)| = o_p(q/\sqrt{T})$ by Theorems 2, definition of $N_T(\zeta_\mu(\theta), \mu)$, and $J_W\left(\theta(\mu)\right)$ being uniformly nonsingular, from which the required conclusion follows.

□

8. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4.0 and ChatGPT 03 to suggest edits to some lengthy passages for concision and clarity. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401, 2001.

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

Isaiah Andrews, Harvey Barnhard, and Jacob Carlson. True and pseudo-true parameters. 2024. URL `https://economics.mit.edu/sites/default/files/2024-04/TrueandPseudo-TrueParameters.pdf`.

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

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

Ivan A. Canay and Azeem M. Shaikh. *Practical and Theoretical Advances in Inference for Partially Identified Models*, page 271–306. Econometric Society Monographs. Cambridge University Press, 2017.

Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*. Springer Science & Business Media, 2004.

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. On uniform asymptotic risk of averaging gmm estimators. *Quantitative Economics*, 10:931–979, 2019.

Victor Chernozhukov and Christian Hansen. The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *The Review of Economics and*

Statistics, 86(3):735–751, 08 2004. ISSN 0034-6535. doi: 10.1162/0034653041811734. URL `https://doi.org/10.1162/0034653041811734`.

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

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

Siddhartha Chib, Minchul Shin, and Anna Simoni. Bayesian estimation and comparison of moment condition models. *Journal of the American Statistical Association*, 113(524):1656–1668, 2018.

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

Uwe Einmahl and Deli Li. Characterization of lil behavior in banach space. *Transactions of the American Mathematical Society*, 360(12):6677–6693, 2008.

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

Christian Francq and Jean-Michel Zakoïan. A central limit theorem for mixing triangular arrays of variables whose dependence is allowed to grow with the sample size. *Econometric Theory*, 21(6):1165–1171, 2005.

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.

A Ronald Gallant and Han Hong. A statistical inquiry into the plausibility of recursive utility. *Journal of Financial Econometrics*, 5(4):523–559, 2007.

Raffaella Giacomini and Toru Kitagawa. Robust bayesian inference for set-identified models. *Econometrica*, 89(4):1519–1556, 2021.

Raffaella Giacomini, Toru Kitagawa, and Matthew Read. Robust bayesian inference in proxy svars. *Journal of Econometrics*, 228(1):107–126, 2022.

Paul Gustafson. *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*. Chapman & Hall, 04 2015. ISBN 9780429192289. doi: 10.1201/b18308.

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.

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

Yuichi Kitamura, Taisuke Otsu, and Kirill Evdokimov. Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica*, 81(3):1185–1201, 2013.

Brendan Kline and Elie Tamer. Bayesian inference in a class of partially identified models. *Quantitative Economics*, 7(2):329–366, 2016.

Yizhou Kuang. Bayesian sensitivity analysis for structural models. 2024.

Yuan Liao and Anna Simoni. Bayesian inference for partially identified smooth convex models. *Journal of Econometrics*, 211(2):338–360, 2019.

Francesca Molinari. Chapter 5 - microeconometrics with partial identification. In Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin, editors, *Handbook of Econometrics, Volume 7A*, volume 7 of *Handbook of Econometrics*, pages 355–486. Elsevier, 2020. doi: https://doi.org/10.1016/bs.hoe.2020.05.002. URL `https://www.sciencedirect.com/science/article/pii/S1573441220300027`.

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

Whitney K Newey and Richard J Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

Andriy Norets and Xun Tang. Semiparametric inference in dynamic binary choice models. *Review of Economic Studies*, 81(3):1229–1262, 2014.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387212396.

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

Vladimir Spokoiny and Maxim Panov. Accuracy of Gaussian approximation in nonparametric Bernstein–von Mises theorem. *arXiv preprint arXiv:1910.06028*, 2019.

James H Stock and Jonathan H Wright. GMM with weak identification. *Econometrica*, 68(5): 1055–1096, 2000.

Plausible GMM: A Quasi-Bayesian Approach

Supplementary Appendix

Victor Chernozhukov[*], Christian B. Hansen[†], Lingwei Kong[‡], Weining Wang[§]

June 30, 2025

[*]Department of Economics, Massachusetts Institute of Technology
[†]Booth School of Business, University of Chicago
[‡]Faculty of Economics and Business, University of Groningen
[§]Department of Economics, University of Bristol

SA-1. VERIFICATION OF ASSUMPTION 9.

We impose a high-level expansion assumption in Assumption 9. We extend the method of obtaining an expansion of a CUE estimator from Newey and Smith (2004) to our high-dimensional misspecified moment settings to verify that assumption. Corollary SA-1 delivers this expansion.

Corollary SA-1. Suppose that the following assumptions hold with $\xi > 2$:

(1) Assumption 1;
(2) Assumption 6;
(3) $g(Z_t,\theta)$ is continuous in $\theta \in \Theta$ with probability one;
(4) for a given $\mu \in \mathscr{M}$, $\sup_{1\le u\le q,\theta\in\Theta}\mathbb{E}(\|e_u^\top(g(Z_t,\theta)-\mu)\|^\xi) < C$, $\sup_{\theta\in\Theta}\mathbb{E}(\|(g(Z_t,\theta)-\mu)(g(Z_t,\theta)-\mu)^\top\|^\xi) < Cq^{\xi/2}$ for some positive constant C, where e_u denotes a $q\times 1$ vector with the uth element of e_u being one and the rest entries are zero;
(5) Define $\|.\|_F$ as the Fubinius norm. For each given $\mu \in \Gamma$, $\mathbb{E}\|(g(Z_t,\theta(\mu))-\mu)(g(Z_t,\theta(\mu))-\mu)^\top-\Omega(\theta(\mu),\mu)\|_F^2 = O(q)$, $\mathbb{E}\|g(Z_t,\theta(\mu))-\mu\|^\xi \lesssim q^{\xi/2}$, $\mathbb{E}\|(g(Z_t,\theta(\mu))-\mu)(g(Z_t,\theta(\mu))-\mu)^\top-\Omega(\theta(\mu),\mu)\|_F^\xi \lesssim q^{\xi/2}$;
(6) $g(Z_t,\theta)$ is continuously differentiable, and $G_t(\theta) = \partial g_t(\theta)/\partial\theta$;
(7) there exists a neighborhood, $\mathcal{N}(\theta(\mu))$, of $\theta(\mu)$ such that for each $\theta \in \mathcal{N}(\theta(\mu))$: there exists a non-negative scalar random variable $b(Z_t)$ such that $\|g(Z_t,\theta)-g(Z_t,\theta(\mu))\| \le b(Z_t)\|\theta-\theta(\mu)\|$, $\|\partial(g(Z_t,\theta)-\mu)/\partial\theta-\partial(g(Z_t,\theta(\mu))-\mu)/\partial\theta\| \le b(Z_t)\|\theta-\theta(\mu)\|$, and $\mathbb{E}[(b(Z_t))^2] < \infty$;
(8) $\|\sup_{\theta\in\Theta}(T^{-1}\sum_t\{G_t(\theta)-\mathbb{E}G_t(\theta)\})\| = o_p(1)$, there exists $C > 0$ such that $1/C \le \lambda_{\min}(\mathbb{E}[(\partial(g(Z_t,\theta)-\mu)/\partial\theta^\top)]^\top\mathbb{E}[\partial(g(Z_t,\theta)-\mu)/\partial\theta^\top]) \le \lambda_{\max}(\mathbb{E}(\partial(g(Z_t,\theta)-\mu)/\partial\theta^\top)^\top\mathbb{E}\partial(g(Z_t,\theta)-\mu)/\partial\theta^\top) \le C$;
(9) There exits a constant $\xi_\lambda \in (\xi^{-1},2^{-1})$ such that $\max\{qT^{-\xi_\lambda+\xi^{-1}}, qT^{\xi_\lambda-2^{-1}}\} \to 0$.

Then, the following expansion (SA-1) holds for any fixed $\mu \in \Gamma$,

$$\|\widehat{\theta}(\mu)-\theta(\mu)-(G(\theta(\mu))^\top\Omega(\theta(\mu),\mu)^{-1}G(\theta(\mu)))^{-1}G(\theta(\mu))^\top\Omega(\theta(\mu),\mu)^{-1}\left(\widehat{m}\left(\theta(\mu)\right)-\mu\right)\| = o_p(qT^{-1/2}), \tag{SA-1}$$

where we consider $\widehat{\theta}(\mu) = \arg\min_{\theta\in\Theta}\{\widehat{m}(\theta)-\mu\}^\top\widehat{\Omega}(\theta,\mu)^{-1}\{\widehat{m}(\theta)-\mu\}$, $\Omega(\theta(\mu),\mu) = \mathbb{E}[(g(Z_t,\theta(\mu))-\mu)(g(Z_t,\theta(\mu))-\mu)^\top]$ and $\widehat{\Omega}(\theta(\mu),\mu) = T^{-1}\sum_{t=1}^T[(g(Z_t,\theta(\mu))-\mu)(g(Z_t,\theta(\mu))-\mu)^\top]$. Thus, Assumption 9 is satisfied for this case once the distributional property of $(\widehat{m}(\theta)-\mu)$ leads to the Gaussian limiting distribution, which hold under conditions indicated by, e.g., the central limit theorem proposed in Francq and Zakoïan (2005).

Before the discussion, we first claim several lemmas employed in proving Corollary SA-1, which correspond to Lemmas A1-A3 in Newey and Smith (2004) and are adjusted to fit our conditions. Furthermore, we let $\rho(v)$ be a quadratic scalar function such that $\rho(v) = -v-\frac{1}{2}v^2$ and let $\rho_j(v) = \partial^j\rho(v)/\partial v^j$, and we let

$$\widehat{P}_\mu(\theta,\lambda) = \frac{1}{T}\sum_{t=1}^T\rho(\lambda^\top(g(Z_t,\theta)-\mu)) = -\lambda^\top\frac{1}{T}\sum_{t=1}^T(g(Z_t,\theta)-\mu)-\lambda^\top\frac{1}{T}\sum_{t=1}^T(g(Z_t,\theta)-\mu))(g(Z_t,\theta)-\mu))^\top\lambda.$$

The estimator for $\theta(\mu)$ for a given $\mu \in \Gamma$ then corresponds to a saddle point problem such that

$$\widehat{\theta}(\mu) = \arg\min_{\theta\in\Theta} \max_{\lambda\in\widehat{\Lambda}_{T,\mu}(\theta)} \widehat{P}_\mu(\theta,\lambda),$$

with $\widehat{\Lambda}_{T,\mu}(\theta) = \{\lambda : \lambda^\top(g(Z_t,\theta)-\mu) \in \Lambda_g, 1 \le t \le T\}$ and Λ_g being an open interval containing zero. We use the following notations for a given $\mu \in \Gamma$. We denote $g_t(\theta) = g(Z_t,\theta)$ and $\widehat{g}_{\mu,t} = g_t(\widehat{\theta}(\mu))$. Let $\upsilon_{\theta\lambda}$ be a augmented parameter vector such that $\upsilon_{\theta\lambda} = (\theta^\top,\lambda^\top)^\top$ and for a given $\mu \in \Gamma$, $\upsilon_{\theta(\mu)0} = (\theta(\mu)^\top, 0^\top)^\top$. Let $v_t(\mu,\upsilon_{\theta\lambda}) = \lambda^\top(g_t(\theta)-\mu)$ and $v_{t,1}(\mu,\upsilon_{\theta\lambda}) = \partial v_t(\mu,\upsilon_{\theta\lambda})/\partial\upsilon_{\theta\lambda} = \left(\lambda^\top G_t(\theta), (g_t(\theta)-\mu)^\top\right)^\top$. Denote $m_\mu(Z_t,\upsilon_{\theta\lambda}) = \partial\rho\left(v_t(\mu,\upsilon_{\theta\lambda})\right)/\partial\upsilon_{\theta\lambda} = \rho_1\left(v_t(\mu,\upsilon_{\theta\lambda})\right)v_{t,1}(\mu,\upsilon_{\theta\lambda})$, $m_{\mu,t}(\upsilon_{\theta\lambda}) = m_\mu(Z_t,\upsilon_{\theta\lambda})$, $M_\mu = \mathbb{E}\left[\partial m_{\mu,t}\left(\upsilon_{\theta(\mu)0}\right)/\partial\upsilon_{\theta\lambda}^\top\right]$, $\psi_\mu(Z_t) = -M_\mu^{-1} m_\mu(Z_t,\upsilon_{\theta\lambda})$ and $\widehat{\psi}_\mu = \sum_{t=1}^T \psi_\mu(Z_t)/\sqrt{T}$. By construction, we have $M_\mu = -\begin{pmatrix} 0 & G(\theta(\mu))^\top \\ G(\theta(\mu)) & \Omega(\theta(\mu),\mu)\end{pmatrix}$.

Lemma SA-11. (Generalised High dimensional Fuk-Nagaev Inequality) Consider i.i.d. centered $X_1,\dots,X_n$ in $\mathbb{R}^k$. Let $\Sigma := \mathbb{E}[X_1X_1^\top]$ and $\omega := \mathrm{diag}(\Sigma)$. Assume that $\mathbb{E}\|X_1\|^\xi < \infty$ for some $\xi > 2$, then we have for any $t > 0$,

$$\mathrm{P}\Big\{\Big\|\sum_{i=1}^n X_i\Big\| \ge 2\sqrt{n|\omega|_1} + t\Big\} \le C_\xi \frac{n\mathrm{E}\|X_1\|^\xi}{t^\xi} + \exp\Big(-\frac{t^2}{3n\|\omega\|}\Big), \tag{SA-2}$$

where $C_\xi > 0$ is a constant only depending on ξ.

Proof. See Theorem 3.1 in Einmahl and Li (2008). We apply Theorem 3.1 therein with $(B,\|\cdot\|) = \left(\mathbb{R}^k,\|\cdot\|\right)$ where $\eta = \delta = 1$. The unit ball of the dual of $\left(\mathbb{R}^k,\|\cdot\|\right)$ is the set of linear functions $\left\{x = (x_1,\dots,x_k)^T \mapsto \sum_{j=1}^k \lambda_j x_j : (\sum_{j=1}^k |\lambda_j|^2)^{1/2} \le 1\right\}$, and for $\lambda_1,\dots,\lambda_k$ with $(\sum_{j=1}^k |\lambda_j|^2)^{1/2} \le 1$, with the following step,

$$\sum_{i=1}^n \mathbb{E}\left[\left(\sum_{j=1}^k \lambda_j X_{ij}\right)^2\right] = \sum_{i=1}^n \mathbb{E}\left[\lambda^\top X_i X_i^\top \lambda\right] \le n\|\lambda\|\lambda_{\max}(\Sigma) \le n\|\omega\|.$$

Hence in this case, Λ_n^2 in Theorem 3.1 Einmahl and Li (2008) is bounded by $n\|\omega\|$. Additionally, by Jensen's inequality, $\sqrt{n|\omega|_1} \ge \mathbb{E}\|\sum_{i=1}^n X_i\|$. Therefore, Lemma SA-11 is implied by Theorem 3.1 Einmahl and Li (2008). ☐

Lemma SA-12. Under the assumptions of Corollary SA-1, let $b_t = \sup_{\theta\in\Theta}\|g(Z_t,\theta)-\mu\|$ for a $\mu \in \Gamma$, then $\max_{1\le t\le T} b_t = O_p(q^{1/2}T^{1/\xi})$.

Proof of Lemma SA-12.
Denote $b_{tu} = \sup_{\theta\in\Theta}\|e_u^\top(g(Z_t,\theta)-\mu)\|$, and thus $b_t \le \max_{1\le u\le q}\sqrt{q}b_{tu}$. The assumption that $\max_{1\le u\le q}\mathbb{E}(b_{tu}^\xi) < C$ and the Markov inequality imply that $\max_{1\le u\le q}\max_{1\le t\le T} b_{tu} = O_p(T^{1/\xi})$, and thus the conclusion follows. ☐

Lemma SA-13. Under the assumptions of Corollary SA-1, $\sup_{\theta\in\Theta,\lambda\in\Lambda_{T,s_\lambda},1\le t\le T}\|\lambda^\top(g(Z_t,\theta)-\mu)\| = O_p(s_\lambda q^{\frac{1}{2}} T^{-\xi_\lambda+\frac{1}{\xi}})$ with $\Lambda_{T,s_\lambda} = \{\lambda : \|\lambda\| \lesssim s_\lambda T^{-\xi_\lambda}\}$ and $s_\lambda > 0$.

Proof of Lemma SA-13. This is a direct result of Lemma SA-12 and Cauchy-Schwarz inequality such that $\sup_{\theta\in\Theta,\lambda\in\Lambda_{T,s_\lambda},1\le t\le T}\|\lambda^\top(g(Z_t,\theta)-\mu)\|\le s_\lambda T^{-\xi_\lambda}\max_{1\le t\le T}b_t = s_\lambda T^{-\xi_\lambda}O_p(q^{\frac{1}{2}}T^{\frac{1}{\xi}})$. $\square$

Lemma SA-14. Under the assumptions of Corollary SA-1, $\bar\lambda=\arg\max_{\lambda\in\Lambda_{T,q^{1/2}}}\widehat{P}_\mu(\theta(\mu),\lambda)$ exists with probability approaching one such that $\bar\lambda=O_p(q^{1/2}T^{-1/2})$ and $\max_{\lambda\in\Lambda_{T,q^{1/2}}}\widehat{P}_\mu(\theta(\mu),\lambda)=O_p(qT^{-1})$ with $\Lambda_{T,q^{1/2}}=\{\lambda:\|\lambda\|\lesssim q^{1/2}T^{-\xi_\lambda}\}$ as in Lemma SA-13.

Proof of Lemma SA-14. $\|\frac{1}{T}\sum_t(g(Z_t,\theta(\mu))-\mu)(g(Z_t,\theta(\mu))-\mu)^\top-\Omega(\theta(\mu),\mu)\|$ $=O_p(q^{1/2}T^{-1/2})=o_p(1)$ by (9) results from Lemma SA-11 and the assumptions $\mathbb{E}\|(g(Z_t,\theta(\mu))-\mu)(g(Z_t,\theta(\mu))-\mu)^\top-\Omega(\theta(\mu),\mu)\|_F^2=O(q)$, $\mathbb{E}\|(g(Z_t,\theta(\mu))-\mu)(g(Z_t,\theta(\mu))-\mu)^\top-\Omega(\theta(\mu),\mu)\|_F^\xi\lesssim q^{\xi/2}$. By the global concavity of $\widehat{P}_\mu(\theta(\mu),\lambda)$ we know $\bar\lambda$ exists and a first-order Taylor expansion around zero for $\widehat{P}_\mu(\theta(\mu),\lambda)$ gives,

$$
\begin{aligned}
0=\widehat{P}_\mu(\theta(\mu),0)\le\widehat{P}_\mu(\theta(\mu),\bar\lambda)&=-\bar\lambda^\top(\widehat{m}(\theta(\mu))-\mu)-\frac{1}{2}\bar\lambda^\top(\frac{1}{T}\sum_t(g_t(\theta(\mu))-\mu)^\top(g_t(\theta(\mu))-\mu))\bar\lambda\\
&\le-\bar\lambda^\top(\widehat{m}(\theta(\mu))-\mu)-\frac{1}{2}(\lambda_{\min}(\Omega(\theta(\mu),\mu))+o_p(1))\|\bar\lambda\|^2\lesssim_p\|\bar\lambda\|\|\widehat{m}(\theta(\mu))-\mu\|-C\|\bar\lambda\|^2,
\end{aligned}
$$

where C is a positive constant. We know $\|\bar\lambda\|\le\|\widehat{m}(\theta(\mu))-\mu\|/C=O_p(q^{1/2}T^{-1/2})$, which guarantees and thus $\sup_{\mu\in\Gamma}\max_{\lambda\in\Lambda_{T,q^{1/2}}}\widehat{P}_\mu(\theta(\mu),\lambda)=O_p(qT^{-1})$. $\square$

Lemmas SA-12-SA-14 are intermediate results used for the proof of Lemma SA-15. Lemma SA-14 provides initial results for the objective function evaluated at the true value $\theta(\mu)$, the rate of which serves as an upper bound in order to derive the rate of $\|\widehat{m}(\widehat\theta(\mu))-\mu\|$ stated in Lemma SA-15. The rate of $\|\widehat{m}(\widehat\theta(\mu))-\mu\|$ is instrumental in deducing a rate of $\|\widehat\theta(\mu)-\theta(\mu)\|$ in Lemma SA-16 and thus the consistency result in Lemma SA-17, leading to the eventual asymptotic normality.

Lemma SA-15. Under the assumptions of Corollary SA-1, $\|\widehat{m}(\widehat\theta(\mu))-\mu\|=O_p(qT^{-1/2})$ for any $\mu\in\Gamma$.

Proof of Lemma SA-15. Since we have a saddle point problem,

$$\widehat{P}_\mu(\widehat\theta(\mu),\widehat\lambda_g(\epsilon_T))\le\widehat{P}_\mu(\widehat\theta(\mu),\widehat\lambda)\le\max_{\lambda\in\Lambda_{T,q^{1/2}}}\widehat{P}_\mu(\theta(\mu),\lambda),$$

where $\widehat\lambda_g(\epsilon_T)=-\epsilon_T(\widehat{m}(\widehat\theta(\mu))-\mu)$ and ϵ_T is a positive scalar. We have from Lemma SA-14 that $\widehat{P}_\mu(\widehat\theta(\mu),\widehat\lambda)\le\max_{\lambda\in\Lambda_{T,q^{1/2}}}\widehat{P}_\mu(\theta(\mu),\lambda)=O_p(qT^{-1})$. For the term $\widehat{P}_\mu(\widehat\theta(\mu),\widehat\lambda_g(\epsilon_T))$ we have that

$$\widehat{P}_\mu(\widehat\theta(\mu),\widehat\lambda_g(\epsilon_T))=\epsilon_T\|\widehat{m}(\widehat\theta(\mu))-\mu\|^2-1/2(\widehat\lambda_g(\epsilon_T))^\top(\frac{1}{T}\sum_t(g_t(\widehat\theta(\mu))-\mu)^\top(g_t(\widehat\theta(\mu))-\mu))\widehat\lambda_g(\epsilon_T),$$

and thus the fact that

$$1/2(\widehat\lambda_g(\epsilon_T))^\top(\frac{1}{T}\sum_t(g_t(\widehat\theta(\mu))-\mu)^\top(g_t(\widehat\theta(\mu))-\mu))\widehat\lambda_g(\epsilon_T)\le q(\sum_{t=1}b_t^2/T)\|\widehat\lambda_g(\epsilon_T)\|^2$$

implied by the proof of Lemma SA-12 leads to

$$\epsilon_T\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|^2-O_p(q)\|\widehat{\lambda}_g(\epsilon_T)\|^2=\epsilon_T\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|^2-O_p(q)\epsilon_T^2\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|^2\leq O_p(qT^{-1})$$

If we choose $\epsilon_T=T^{-\xi_\lambda}/\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|$ so that $\widehat{\lambda}_g(T^{-\xi_\lambda}/\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|)\in\Lambda_{T,q^{1/2}}\cap\widehat{\Lambda}_T(\widehat{\theta}(\mu))$ w.p.a.1., the above inequality implies that

$$\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|\leq O_p(q)O_p(T^{-\xi_\lambda})+O_p(qT^{\xi_\lambda-1}).$$

Hence, $\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|=o_p(T^{-1/\xi})$ and from the assumption $qT^{-\xi_\lambda+\xi^{-1}}\to0$, we know that $\widehat{\lambda}_g(\epsilon_T)$ is an interior point in $\Lambda_{T,q^{1/2}}$ and in $\widehat{\Lambda}_T(\widehat{\theta}(\mu))$ w.p.a.1. as long as $\epsilon_T\to0$. Now consider one arbitrary drifting to zero sequence ϵ_T, $\epsilon_T\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|^2-O_p(q)\epsilon_T^2\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|^2=O_p(qT^{-1})$ implies that $\epsilon_T\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|^2=O_p(qT^{-1})$ if $\epsilon_T=o_p(q^{-1})$, since this is true for all such sequences, we know $\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|=O_p(qT^{-1/2})$. □

Lemma SA-16. Under the assumptions of Corollary SA-1, $\|\widehat{\theta}(\mu)-\theta(\mu)\|=O_p(qT^{-1/2})$.

Proof of Lemma SA-16. The differentiability assumption implies that

$$\widehat{m}(\widehat{\theta}(\mu))-\widehat{m}(\theta(\mu))=T^{-1}\sum_t G_t(\bar{\theta}(\mu))^\top(\widehat{\theta}(\mu)-\theta(\mu)),$$

where

$$\|\bar{\theta}(\mu)-\theta(\mu)\|\leq\|\widehat{\theta}(\mu)-\theta(\mu)\|.$$

Denote $G(\theta)=\mathbb{E}G_t(\theta)$, and the assumption $\|\sup_{\theta\in\Theta}(T^{-1}\sum_t G_t(\theta)^\top-\mathbb{E}G_t(\theta)^\top)\|=o_p(1)$ leads to $\widehat{m}(\widehat{\theta}(\mu))-\widehat{m}(\theta(\mu))=(G(\bar{\theta}(\mu))+o_p(1))^\top(\widehat{\theta}(\mu)-\theta(\mu))$, where $o_p(1)$ is defined in terms of $|.|_2$ norm. Thus the assumption that $\lambda_{\min}(G(\bar{\theta}(\mu))^\top G(\bar{\theta}(\mu)))$ is bounded away from zero implies that

$$\|(\widehat{\theta}(\mu)-\theta(\mu))\|\leq C\|\widehat{m}(\widehat{\theta}(\mu))-\widehat{m}(\theta(\mu))\|\leq C\{\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|+\|\widehat{m}(\theta(\mu))-\mu\|\},$$

then the rate follows as $\|\widehat{m}(\widehat{\theta}(\mu))-\mu\|=O_p(qT^{-1/2})$ indicated by Lemma SA-15 and $\|\widehat{m}(\theta(\mu))-\mu\|=O_p(q^{1/2}T^{-1/2})$ indicated by Lemma SA-11. □

Consider a first-order Taylor expansion of $\sum_t m(Z_t,\upsilon_{\theta\lambda})/T$ for $\widehat{\upsilon}_{\theta\lambda}=\left(\widehat{\theta}(\mu)^\top,\widehat{\lambda}^\top\right)^\top$ and $\upsilon_{\theta(\mu)0}=\left(\theta(\mu)^\top,0^\top\right)^\top$, we have $0=\begin{pmatrix}0\\-\widehat{m}\left(\theta(\mu)\right)\end{pmatrix}+\bar{M}_\mu\left(\widehat{\upsilon}_{\theta\lambda}-\upsilon_{\theta(\mu)0}\right)$, where

$$\bar{M}_\mu=\begin{pmatrix}0 & \sum_{t=1}^T\rho_1\left(\bar{\lambda}^\top\widehat{g}_{\mu,t}\right)G_t(\bar{\theta}(\mu))^\top/T\\ \sum_{t=1}^T\rho_1\left(\bar{\lambda}^\top\widehat{g}_{\mu,t}\right)G_t(\bar{\theta}(\mu))/T & \sum_{t=1}^T\rho_2\left(\bar{\lambda}^\top\widehat{g}_{\mu,t}\right)g_t(\bar{\theta}(\mu))\widehat{g}_{\mu,t}^\top/T\end{pmatrix},$$

and $\bar{\theta}(\mu)$ and $\bar{\lambda}$ are mean values converging to $\theta(\mu)$ and 0 that actually differ from row to row of the matrix $\bar{M}_\mu$.

Lemma SA-17. Under the assumptions of Corollary SA-1, $\|\bar{M}_\mu-M_\mu\|=o_p(1)$, for all μ.

Proof of Lemma SA-17. We first show that $\|\widehat{\Omega}(\theta(\mu),\mu)-\Omega(\theta(\mu),\mu)\| = o_p(1)$ with $\widehat{\Omega}(\theta(\mu),\mu) = \frac{1}{T}\sum_t(\widehat{g}_{\mu,t}-\mu)(\widehat{g}_{\mu,t}-\mu)^\top$. This comes directly from the fact that

$$\begin{aligned}
&\|\widehat{\Omega}(\theta(\mu),\mu)-\Omega(\theta(\mu),\mu)\| = \|\frac{1}{T}\sum_t(\widehat{g}_{\mu,t}-\mu)(\widehat{g}_{\mu,t}-\mu)^\top-\Omega(\theta(\mu),\mu)\|\\
\le& \left\|\frac{1}{T}\sum_t(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))^\top\right\| + 2\left\|\frac{1}{T}\sum_t(g_t(\theta(\mu))-\mu)(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))^\top\right\|\\
&+\left\|\frac{1}{T}\sum_t(g_t(\theta(\mu))-\mu)(g_t(\theta(\mu))-\mu)^\top-\Omega(\theta(\mu),\mu)\right\| = o_p(1)
\end{aligned}$$

where

(1) $\left\|\frac{1}{T}\sum_t(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))^\top\right\| = o_p(1)$ results from the fact that $\|\widehat{g}_{\mu,t}-g_t(\theta(\mu)))\| \le b(Z_t)\|\widehat{\theta}(\mu)-\theta(\mu)\|$ and
$$\left\|\frac{1}{T}\sum_t(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))^\top\right\| \le \frac{1}{T}\sum_t(b(Z_t))^2\|\widehat{\theta}(\mu)-\theta(\mu)\|^2.$$

(2) $\left\|\frac{1}{T}\sum_t(g_t(\theta(\mu))-\mu)(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))^\top\right\| = o_p(1)$ results from the fact that
$$\begin{aligned}
&\left\|\frac{1}{T}\sum_t(g_t(\theta(\mu))-\mu)(\widehat{g}_{\mu,t}-g_t(\theta(\mu)))^\top\right\|\\
=&\left\|\frac{1}{T}\sum_t(g_t(\theta(\mu))-\mu)(\widehat{\theta}(\mu)-\theta(\mu)))^\top G_t(\bar{\theta}(\mu))^\top\right\|\\
\le&\max_t b_t\|\widehat{\theta}(\mu)-\theta(\mu))\|\left\|\frac{1}{T}\sum_t G_t(\bar{\theta}(\mu))^\top\right\| = O_p(q^{3/2}T^{1/\xi-1/2})
\end{aligned}$$

(3) $\left\|\frac{1}{T}\sum_t(g_t(\theta(\mu))-\mu)(g_t(\theta(\mu))-\mu)^\top-\Omega(\theta(\mu),\mu)\right\| = O_p(q^{1/2}T^{-1/2})$ as in the proof of Lemma SA-14.

Next, $\|\frac{1}{T}\sum_t G_t(\widehat{\theta}(\mu))-\mathbb{E}G_t(\theta(\mu)))\| = o_p(1)$ results from the triangular inequality and the facts that
$$\|\frac{1}{T}\sum_t G_t(\widehat{\theta}(\mu))-G_t(\theta(\mu)))\| \le \frac{\sum_t b(Z_t)}{T}\|\widehat{\theta}(\mu)-\theta(\mu)\|,$$
and that $\|\frac{1}{T}\sum_t G_t(\theta(\mu))-\mathbb{E}G_t(\theta(\mu)))\|$ is bounded by assumption. Since the above arguments also hold when we replace parts of $\widehat{\theta}(\mu)$ by $\bar{\theta}(\mu)$, the final conclusion is then indicated once we notice that $\|\bar{\lambda}^\top\widehat{g}_{\mu,t}\| = o_p(1)$, which is indicated by Lemma SA-15 and the fact that $\|\widehat{\Omega}-\Omega(\theta(\mu),\mu)\| = o_p(1)$.

□

Proof of Corollary SA-1. We now can establish the expansion SA-1 to show the validity of Assumption 9. Note that by assumption M_μ is non-singular, and Lemma SA-17 implies that $\bar{M}_\mu$ is also non-singular w.p.a.1. such that $\|\bar{M}_\mu^{-1}-M_\mu^{-1}\| \le \|\bar{M}_\mu^{-1}\|\|\bar{M}_\mu-M_\mu\|\|M_\mu^{-1}\| = o_p(1)$, and thus

$$\sqrt{T}\left(\widehat{v}_{\theta\lambda}-v_{\theta(\mu)0}\right) = -(M_\mu^{-1}+o_p(1))\left(0,-\sqrt{T}(\widehat{m}\left(\theta(\mu)\right)^\top-\mu^\top)\right)^\top,$$

with

$$M_\mu^{-1} = -\begin{pmatrix} F_\mu^{-1} & -F_\mu^{-1}G(\theta(\mu))^\top\Omega(\theta(\mu),\mu)^{-1} \\ -\Omega(\theta(\mu),\mu)^{-1}G(\theta(\mu))F_\mu^{-1} & \Omega(\theta(\mu),\mu)^{-1}+\Omega(\theta(\mu),\mu)^{-1}G(\theta(\mu))F_\mu^{-1}G(\theta(\mu))^\top\Omega(\theta(\mu),\mu)^{-1} \end{pmatrix}$$

and $F_\mu = -G(\theta(\mu))^\top\Omega(\theta(\mu),\mu)^{-1}G(\theta(\mu))$. Therefore, $\|(\widehat{\theta}(\mu)-\theta(\mu))-F_\mu^{-1}G(\theta(\mu))^\top\Omega(\theta(\mu),\mu)^{-1}(\widehat{m}\left(\theta(\mu)\right)-\mu)\| = o_p(qT^{-1/2})$. The assumptions proposed in this analysis can be relaxed further if only for the validity of Assumption 9. □

SA-2. LINK TO OPTIMAL DECISION RULE

We argue in this section that, similar to the results established in Andrews and Mikusheva (2022), under model misspecification, we can also establish that the quasi-posterior can be obtained as the limit of a sequence of posteriors under proper priors, and the resulting quasi-Bayes decision rule can correspond to the pointwise limit of the sequence of Bayes decision rules. To align with the setup in Andrews and Mikusheva (2022), we do not work directly with $\mathbb{P}_\mu$ or focus solely on the parameter $\theta(\mu)$. Instead, We define a fixed coarse reference measurement $\mathbb{P}_0$ and consider the parameter pair $(\theta(\mu),\mu)$. While the construction in this subsection may appear tedious, it is primarily introduced to facilitate this match.

Assumption SA-12. *(Measure in the limit) There exists $\tilde{Z}_t$, which is a maximal subset of Z_t such that the distribution of $\tilde{Z}_t$ does not change with respect to any $(\theta(\mu),\mu)\in\Xi$. There exists i.i.d. random sequence μ_t such that $\mathbb{E}_{\mathbb{P}_{\mu^*}}(\mu_t)=\mu^*$. The random variable $(g(Z_t,\theta(\mu))-\mu_t,\tilde{Z}_t)$, $t=1,\cdots,T$ follows $\mathbb{P}_0$ that is invariant for all plausible pairs $(\theta(\mu),\mu)\in\Xi$.*

The specification of the $\mathbb{P}_0$ matches the nonparametric Bayesian idea from Gallant and Hong (2007) such that we can treat μ_t as a latent variable capturing the misspecification, and $\mathbb{P}_0$ remains invariant with respect to the plausible characteristics. While $\mathbb{P}_\mu$ is a marginal measure over Z_t given μ, $\mathbb{P}_0$ can then be seen as a coarse joint measure concerning (Z_t,μ_t) and it may not be enough to deduce a marginal distribution for Z_t or μ_t, which is similar to the case discussed in Gallant and Hong (2007).

Assumption SA-12 essentially assumes that the measure (in the limit) of the sample moment can be decomposed into two parts. The plausible term impacts one part, while the remaining part is a measure invariant to $(\theta(\mu),\mu)$. For example, in the plausible IV model from Conley et al. (2012), we may choose $\mu_t=\gamma D_t^\top D_t$, $\tilde{Z}_t=(X_t^\top,W_t^\top,D_t^\top)^\top$ and thus the distribution of $\sum_t[g(Z_t,\theta(\mu))-\mu_t]$ does not depend on the plausible term μ. Additionally, it is reasonable to consider the deviations from the measure of $((g(Z_t,\theta)-\mu_t)^\top,\tilde{Z}_t^\top)^\top$, i.e., $\mathbb{P}_0$, and in the spirit of Andrews and Mikusheva (2022) and Kitamura et al. (2013) we consider perturbations in the probability measure in Assumption SA-13. The subspace of score functions as

$$T_\mu(\mathbb{P}_0)=\left\{f\in T(\mathbb{P}_0):\mathbb{E}_{\mathbb{P}_0}\left[f(Z_t)\left(g\left(Z_t,\theta(\mu)\right)-\mu_t\right)\right]=0\right\}.$$

We define $\bar{m}(\mu,\theta)=\mathbb{E}_{\mathbb{P}_0}(f\left(g\left(Z_t,\theta\right)-\mu_t\right))$, then for $f\in T_\mu(\mathbb{P}_0)$, by design $\bar{m}(\mu,\theta(\mu))=0$.

Assumption SA-13. *(Differentiability in quadratic mean)* $(g(Z_t,\theta)-\mu_t,\tilde{Z}_t)$ *of size* T *follows distribution* $\mathbb{P}=\mathbb{P}_{T,f}$*, where the sequence* $\mathbb{P}_{T,f}$ *converges to* $\mathbb{P}_0$*,*

$$\int\left[\sqrt{T}\left(d\mathbb{P}_{T,f}^{1/2}-d\mathbb{P}_0^{1/2}\right)-\frac{1}{2}f\,d\mathbb{P}_0^{1/2}\right]^2\to 0.$$

While the score f controls the data distribution, our interest lies in the plausible pair $(\theta(\mu),\mu)$. Denote $\widehat{g}(\theta,\mu)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[g(Z_t,\theta)-\mu]$.

Assumption SA-14. *(GP in the limit) Assume that under* $\mathbb{P}_{T,f}$*,* $\widehat{g}(\theta(\mu),\mu)$ *weakly converges to a Gaussian process with mean function* $\bar{m}(\cdot)$ *and covariance function* $\Sigma(\cdot)$*, where the covariance function is continuous and nonsingular for all pairs* (θ,μ) *and is consistently estimable.*

Additionally, denote by $\tilde{g}(\cdot)$ the limiting Gaussian process of $\widehat{g}(\cdot)$ on $(\theta(\mu),\mu)$ pairs, $\mathscr{C}_\Xi$ the space of continuous functions from Ξ to $\mathbb{R}^q$. Let A be any linear functional $A:\mathscr{C}_\Xi\to\mathbb{R}^q$ such that $\mathrm{Cov}(A(\tilde{g}(\cdot)),\tilde{g}(\theta,\mu))$ is nonsingular for all (θ,μ). Let $h_\perp(\cdot)=\tilde{g}(\cdot)-\mathrm{Cov}(\tilde{g}(\cdot),\xi_A)\mathrm{Var}(\xi_A)^{-1}\xi_A$ and $\mu_\perp(\cdot)=\bar{m}(\cdot)-\mathrm{Cov}(\tilde{g}(\cdot),\xi_A)\mathrm{Var}(\xi_A)^{-1}A(\tilde{g}(\cdot))$, where $\xi_A=A(\tilde{g}(\cdot))$.

Analogue to arguments in Sections 3.1.2 from Andrews and Mikusheva (2022), the likelihood function $\ell\left(\mu_\perp,\theta(\mu),\mu;\tilde{g}(\cdot)\right)$ based on the observed data $\tilde{g}(\cdot)$ factors as

$$\ell\left(\mu_\perp,\theta(\mu),\mu;\tilde{g}(\cdot)\right)=\ell\left(\mu_\perp,\theta(\mu),\mu;\xi_A\right)\ell(\mu_\perp;h_\perp),$$

and thus their equation (8) directly implies that a class of proportional priors over $\mu_\perp$ would give rise to quasi-Bayes as a limiting case.

To establish that quasi-Bayes decision rules are pointwise limits of the corresponding Bayes decision rules, it suffices to verify the conditions of Theorem 8 in Andrews and Mikusheva (2022). The non-singularity of the covariance function is ensured, for instance, by Assumption 3.i). The remaining conditions of their Theorem 8 require a loss function $\ell(a,\theta,\mu)$ that is uniformly bounded, continuous, and strictly convex in a for each (θ,μ), along with a compact and convex action space $\mathscr{A}$. These requirements are met with appropriately chosen loss functions and action spaces.

SA-3. Additional Simulation Results

From the Bayesian perspective, Sections SA-3.1.1 and SA-3.2.1 compare PGMM posteriors in various settings and validate our results stated in Lemma 1. From the frequentist perspective, we validate our BvM theorem and related coverage results. In paticular, with the linear IV model setup, we illustrate the validity of Theorems 1-2. In addition to the linear moment case, Section SA-3.2 addresses simulations involving nonlinear moments and illustrates the frequentist justification of the unions of the credible regions constructed via quasi-posterior distributions, as indicated by Theorems 3-4, by evaluating the frequentist coverage rates.

SA-3.1. **Linear moments scenario.**

SA-3.1.1. *Baysian motivation.*
This subsection compares PGMM posteriors between the case where $\mu \equiv 0$ (no model misspecification) and μ draws from F_μ. The posteriors behave differently in these two cases; e.g., the posteriors of the plausible characteristics in the presence of model misspecification tend to deviate from zero instead of centering around zero as priors, which suggests that the shape of the PGMM posterior provides meaningful information about the extent of misspecification.

The simulation exercises in this subsection draw plausible characteristics μ together with Z_t jointly and continue with the linear IV model from Section 3.1, and for simplicity, we consider the case with one endogenous variable, no exogenous control variables, and a single instrument variable. Therefore, $\theta = (\alpha, \beta_X) \in \Theta$ and moment conditions $g(Z_t, \theta) = (1, D_t^\top)^\top \left(Y_t - \alpha + X_t^\top \beta_X\right)$. In this simulation exercise, with a given value of γ, Y_t is generated with X_t, D_t in the following way:

$$Y_t = \alpha + X_t \beta_X + (1, D_t^\top)\gamma + \varepsilon_t,$$

where ε consists of independent and identically distributed (i.i.d.) draws from $N(0, \sigma^2)$, and X_t and D_t are the fixed in the simulation using the sample data as described in Section 3.1. By the fact that $\mu = \mathbb{E}g(Z_t, \theta)$, we have $\gamma = (\mathbb{E}(1, D_t^\top)^\top (1, D_t^\top))^{-1}\mu$, so once we draw μ from F_μ the value of γ is also determined. In the simulation exercises in this subsection, we thus calibrate σ^2 to the data employed in Section 3.1, draw the value of $\mu = (\mu_1, \mu_2)^\top$ from given F_μ, let $\gamma = (\mathbb{E}(1, D_t^\top)^\top (1, D_t^\top))^{-1}\mu$, $\alpha = 0, \beta_X = \mu_2/10$, and then generate Y_t and construct PGMM posteriors based on $Z_t = (Y_t, X_t, D_t)$.

Figure SA-12 plots the results based on the above simulation setup and compares the posteriors in two cases. In the upper panel, μ is fixed at zero, and there is no model misspecification. The lower panel uses DGPs with μ drawn from $\pi(\mu)$; therefore, μ varies instead of fixing at zeros.

In addition, Figure SA-12 shows that the highest posterior regions of β_X tend to cover the true values of β_X used in the DGP. Corresponding to Lemma 1, the average coverage rates of the 95% PR_T containing the values of θ_X used in the DGP are 0.99 and 0.96 for cases (a.) and (b.), respectively.

SA-3.1.2. *Frequentist validation.*
Apart from Section SA-3.1.1 and SA-3.2.1, the remaining sections consider DGPs with fixed plausible characteristics from a frequentist perspective. We continue with the linear model from Section SA-3.1.1. We set α to be zero in this subsection and consider $\theta = \beta_X \in \Theta$. With a given value of γ, Y_t is generated similarly as before:

$$Y_t = X_t \theta + D_t \gamma + \varepsilon_t.$$

In this simulation exercise, we calibrate parameters to the 401(K) data employed in Conley et al. (2012) so that X_t and D_t are fixed subsamples of size 9951 from the 401(K) data with X_t being an

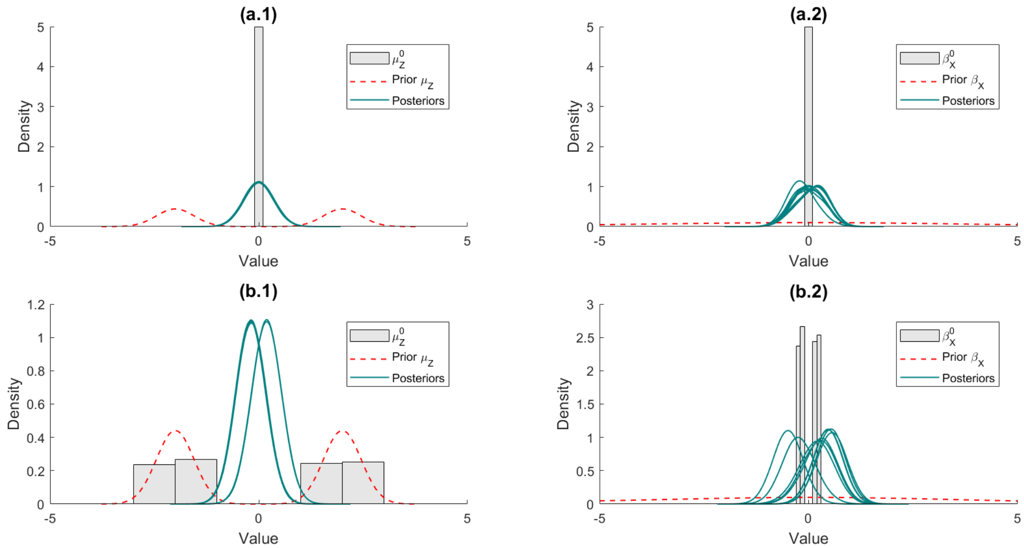

FIGURE SA-12. Marginal priors (red dashed line) and PGMM marginal posteriors (green solid line) of slope coefficient, β_X (left panel), and plausible characteristic, $\mu_Z = \mu_2$ (right panel), resulting from the Section SA-3.1.1 linear IV model simulations. The upper panel (a.) represents a correctly specified model with $\mu \equiv 0$, i.e., F_μ assigns all mass to 0. In contrast, the lower panel (b.) reflects a plausible IV setting from Conley et al. (2012) with μ drawing from the prior of μ, $\pi(\mu)$ in the simulation data generating process (DGP), i.e., F_μ coincides with $\pi(\mu)$. The red dashed curves mark the marginal prior densities. The gray shadowed bars are the histograms of the realized values of β_X, μ used in the DGPs of simulations, i.e., the bar in the subfigure (a.1) locates at point 0 as the value of μ is fixed at 0 in the upper panel simulation exercise. The green solid curves mark the PGMM marginal posteriors, e.g., the green curves in the subfigure (a.1) correspond to the marginal PGMM posteriors of μ_X.

indicator for 401(k) participation and D_t being an indicator for 401(k) plan eligibility. ε consists of independent and identically distributed (i.i.d.) draws from $N(0, \sigma^2)$, and the values of θ and σ^2 are obtained by regressing the net financial assets from the dataset on X_t with D_t as an instrument using the 2SLS estimator and sample variance of the residuals.

SA-3.1.3. *Validity of the Gaussian mixture limiting distribution.*
This section validates the Gaussian mixture limiting distribution specified in Theorem 2 using the linear IV model mentioned above. The simulated data, $Z_t = (Y_t, X_t, D_t), 1 \le t \le T$, is generated as follows: $X_t = D_t + v_t$, $Y_t = X_t\theta + D_t\gamma + \varepsilon_t$, where D_t's are independently log-normally distributed such that the natural logarithm of D_t follows the standard normal distribution, v_t's and ε_t's are independently standard normally distributed, and Z_t, v_t, and ε_t are all independent of each other.

We first proceed with the case $k = q = 1$, and the parameter of interest θ is fixed at zero. We choose $\gamma = 1/\sqrt{T}$ with T being the sample size to mimic local misspecification and use a local Gaussian prior for μ and an independent flat prior for θ such that $\pi(\theta, \mu) \propto \exp(-T\mu^\top \mu/2)$.

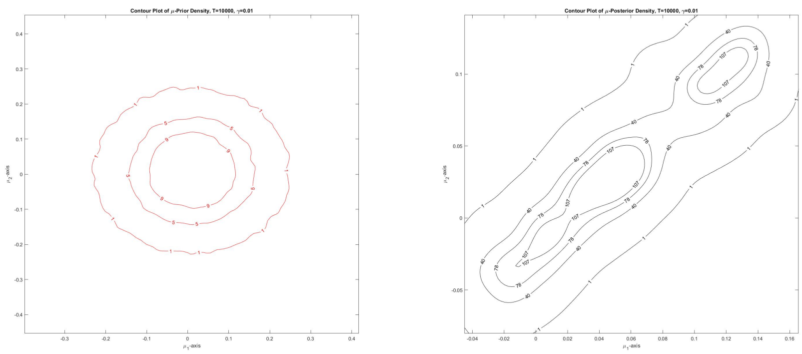

FIGURE SA-13. $q = 2$, with $\gamma = (0.01, 0.01)^{\top}$. Left panel: contour plot of the (marginal) prior density of $\mu = (\mu_1, \mu_2)^{\top}$; Right panel: contour plot of the (marginal) quasi-posterior density of $\mu = (\mu_1, \mu_2)^{\top}$.

Figures SA-14 plots based on random draws of pairs of (θ, μ) from quasi-posterior distributions. These quasi-posterior distributions are constructed using simulated data of various sample sizes. The third and fourth columns of Figure SA-14 illustrate that the Gaussian mixture distribution $N_T(\theta, \mu)$ closely approximates $p_T(\theta, \mu)$ when the sample size is relatively large, as the simulation results show that the conditional quasi-posterior distribution of θ given μ closely resembles a Gaussian. These findings are also shown in the QQ plots of Figure SA-14.

Another interesting observation is that in Figure SA-14, the fifth column indicates that the region with the highest density of quasi-posterior distributions tends to concentrate in a smaller area than the priors. This phenomenon is illustrated in Figure SA-13, where we explore a simulation exercise similar to the settings in Figure SA-14, but with $q = 2$ and $\gamma = (0.01, 0.01)^{\top}$. Figure SA-13 demonstrates that in the presence of over-identification, the posterior distribution may concentrate on an area of smaller dimension than the prior distribution.

SA-3.2. **Nonlinear moments scenario.**

SA-3.2.1. *Baysian motivation.*

We now consider a simulation exercise with non-smooth moment conditions, i.e. the IVQR model. Inspired by the plausible IV model, we use invalid/plausible IV ($\tilde{D}_t$) to introduce model misspecification in this simulation exercise such that $\tilde{D}_t = D_t + \tilde{\gamma}\left(\tau - 1\left(Y_t \leqslant \alpha_\tau + X_t^{\top}\beta_{\tau,X} + W_t^{\top}\beta_{\tau,W}\right)\right)$ with D_t being a valid instrumental variable. Therefore, when $\tilde{\gamma} = 0$ the model is correctly specified, and with non-zero $\tilde{\gamma}$'s, the moment conditions are misspecified. Specifically, this section focuses upon the Median IV case with $\tau = 0.5$ as in Section 3.1 with moment conditions

$$g(Z_t, \theta) = (1, \tilde{D}_t^{\top}, W_t^{\top})^{\top}\left(\tau - 1\left(Y_t \leqslant \alpha_\tau + X_t^{\top}\beta_{\tau,X} + W_t^{\top}\beta_{\tau,W}\right)\right),$$

where $Z_t = (Y_t, X_t, W_t, \tilde{D}_t)$, and $\theta = (\alpha_\tau, \beta_{\tau,X}, \beta_{\tau,W}) \in \Theta$.

Figure SA-15 is created similarly to Figure SA-12 with one instrumental variable. In the simulation exercises in this subsection, X_t, W_t, and D_t are fixed using the sample data as

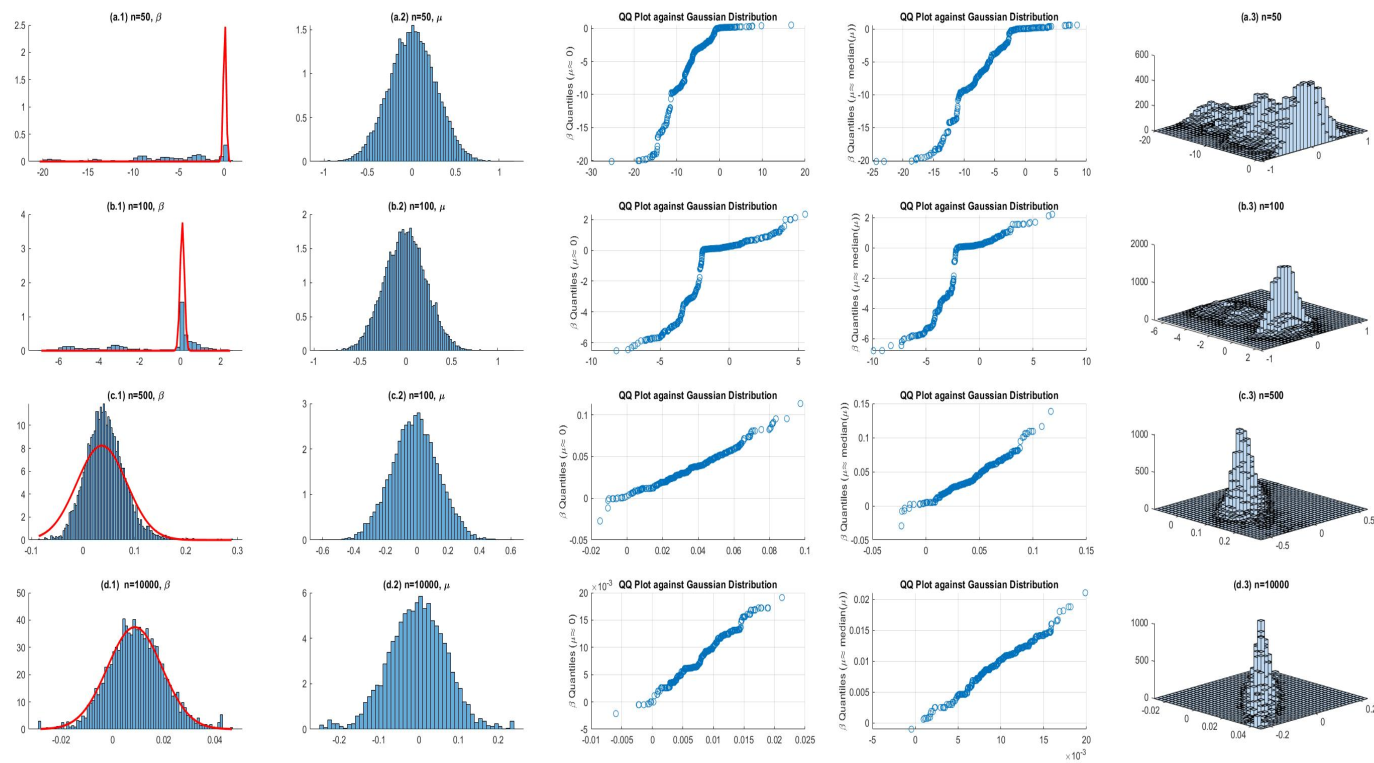

FIGURE SA-14. The first column displays histograms of randomly drawn θ values from quasi-posterior distributions, with the red smooth curve representing the limiting density suggested by Theorem 1. In the second column, you can find histograms of random draws of the plausible characteristic μ from quasi-posterior distributions. The final column exhibits joint histograms of pairs (θ, μ). The third and fourth columns display QQ-plots of simulated θ's corresponding to μ's with values close to zero and the median of the simulated μ's, respectively. The sample size varies from 50 (first row) to 10,000 (fourth row).

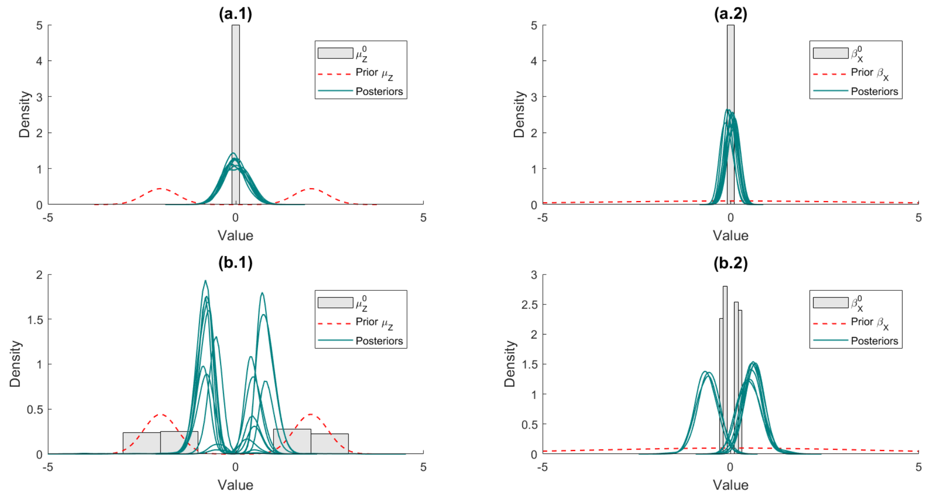

FIGURE SA-15. Similar to Figure SA-12, marginal priors (red dashed line) and PGMM marginal posteriors (green solid line) of slope coefficient, β_X (left panel), and plausible characteristic, μ_D (right panel), resulting from the Section SA-3.2.1 median IV model simulations. The upper panel (a.) represents a correctly specified model with $\mu \equiv 0$. In contrast, the lower panel (b.) uses DGPs with μ drawn from the prior $\pi(\mu)$. The red dashed curves mark the marginal prior densities. The gray shadowed bars are the histograms of the realized values of β_X, μ used in the DGPs of simulations. The green solid curves mark the PGMM marginal posteriors.

described in Section 3.1. We first draw $\mu = (\mu_1, \mu_D, \mu_W)$ from F_μ, and the values of μ directly determines the value of $\tilde{\gamma}$ via $\mu = \mathbb{E}g(Z_t, \theta)$. In this simulation design, μ_1 and μ_W are always set to zero, and once we draw μ, we choose $\theta = \mu/10$, and Y_t is generated in the following way:

$$Y_t = \alpha_\tau + X_t \beta_{\tau,X} + W_t \beta_{\tau,W} + \varepsilon,$$

where ε consists of independent and identically distributed (i.i.d.) draws from $N(0, \sigma^2)$ with σ^2 calibrated to the data employed in Section 3.1.

Similar to Figure SA-12, in the upper panel of Figure SA-15 $\mu \equiv 0$, while in the lower panel, μ used in the DGPs are drawn from the prior $\pi(\mu)$. The observed patterns are similar. The average coverage rates of the 95% PR_T are 0.99 and 0.92 for cases (a.) and (b.) in Figure SA-15, respectively. In the absence of model misspecification in the upper panel of Figure SA-15, the marginal posteriors of the plausible characteristics tend to cluster around zero, and the average coverage rate is undoubtedly higher than the nominal rate due to the additional uncertainty introduced. However, they start to deviate from zero as model misspecification is introduced in the lower panel.

SA-3.2.2. *Frequentist validation.*

This subsection revisits an median IV simulation example from Chernozhukov and Hong (2003) with slight modifications to introduce model misspecification. The Monte Carlo Simulation

γ	T	τ	Methods	$\beta_{\tau,1}$	$\beta_{\tau,2}$	$\beta_{\tau,3}$	γ	T	τ	Methods	$\beta_{\tau,1}$	$\beta_{\tau,2}$	$\beta_{\tau,3}$
1	300	0.2	0	0.946	0.941	0.029	1	300	0.2	1	0.985	0.981	0.924
1	300	0.5	0	0.675	0.676	0.34	1	300	0.5	1	0.988	0.97	0.921
1	100	0.2	0	0.781	0.725	0.273	1	100	0.2	1	0.928	0.929	0.964
1	100	0.5	0	0.554	0.549	0.454	1	100	0.5	1	0.996	0.968	0.94
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	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

TABLE SA-2. This table illustrates the average coverage rates of sets containing the true values of $\beta_{\tau,i}$'s, and the rates are displayed in the columns labeled "$\beta_{\tau,i}$". The column labeled "γ" shows the values of γ used in the DGPs, labeled "T" shows the value of the sample size, labeled "τ" shows the corresponding quantiles, labeled "Methods" outlines the estimation procedure, with the value of 0 referring to the [CH] method and the value of 1, referring to the PGMM method. The [CH] intervals are constructed with flat priors over θ_τ's, and the PGMM intervals are constructed with flat priors for θ_τ's and independent local Gaussian priors $N(0, I/T)$ for plausible characteristics.

Example II established by Chernozhukov and Hong (2003) is:

$$Y_t = \alpha + X_t^\top \beta + u_t, u_t = \sigma(X_t)\varepsilon_t, \quad \sigma(X_t) = (1 + \sum_{j=1}^{3} X_{t,j})/5,$$

where there are no endogenous variables, $X_{t,j}$'s are independently log-normally distributed such that the natural logarithm of $X_{t,j}$ follows a standard normal distribution, ε_i's are independently standard normally distributed and are independent of $X_{t,j}$'s and $\theta = (\alpha, \beta)$. They consider the following moment conditions for the median $g(Z_t, \theta) = (1, X_t^\top)^\top \left(0.5 - 1\left(Y_t \leqslant \alpha + X_t^\top \beta\right)\right)$.

We modify the above DGP to introduce model misspecification in two ways: by adjusting the credibility of the instrument variables and the rank invariance (or similarity) used in the IVQR with discrete (or bounded continuous) treatment variables so that treatment status should not impact the underlying conditional distribution. The former situation resembles the plausible linear IV model in that the exclusion restriction is relaxed. We substitute u_t with $\tilde{u}_t = \sigma(X_t)\varepsilon_t + \gamma X_{t,3}^2$ to include plausible IVs, with γ evaluating the credibility of IVs. In the latter case, we consider the following DGP with potentially missing variables X_t such that $Y_t = \alpha + D_t^\top \beta + \gamma D_t X_t + \varepsilon_t$, where D_t follows an independent and identically distributed Bernoulli distribution with a success rate of $1/2$. For both cases, we consider the following moment conditions used for estimating the parameters concerning the τ-quantile: $g(Z_t, \theta_\tau) = (1, X_t^\top, W_t^\top)^\top \left(\tau - 1\left(Y_t \leqslant \alpha_\tau + X_t^\top \beta_\tau\right)\right)$, where $Z_t = (Y_t, X_t)$, $\theta = (\alpha_\tau, \beta_\tau) \in \Theta$.

Table SA-2 is constructed with simulated data generated under the first DGP relaxing the exclusion restriction, while Table SA-3 is built under the latter one. Both tables compare average coverage rates of sets constructed using different approaches containing true parameter values, and they also demonstrate the validity of Theorems 3 and 4.

γ	T	τ	Methods	β_τ	γ	T	τ	Methods	β_τ
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	1	300	0.5	1	0.9555
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	1	300	0.8	1	0.9009
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	1	100	0.5	1	0.9587
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	1	100	0.8	1	0.9659

TABLE SA-3. Similar to Table SA-2, the value of -1 in the column labeled “Methods” refers to the IVQR method (see Chernozhukov and Hansen (2005)). The [CH] intervals are constructed with flat priors over θ's, and the PGMM intervals are constructed with flat priors for θ's and independent local Gaussian priors $N(0, 10I/T)$ for plausible characteristics.

Table SA-2 compares results with and without accounting for model misspecification. In the simulation exercise outlined in Table SA-2, the parameters $\theta = (\alpha, \beta)$ in the DGP are set equal to the null vector, and thus β_τ's are also null vectors following the settings in Chernozhukov and Hong (2003). The table shows the average coverage rates for the true β_τ value in sets from two quasi-posterior distributions, with and without assuming model misspecification. One set is created using 0.025 to 0.975 sample quantiles obtained from quasi-posterior distributions for $\beta_{\tau,i}$'s without assuming model misspecification (refer to the quasi-Bayesian approach outlined in Chernozhukov and Hong (2003), denoted as [CH]), and another set is constructed from PGMM quasi-posterior. The sets resulting from PGMM approach are built as follows: we select simulated $\beta_{\tau,i}$'s corresponding to pre-selected plausible characteristics $(\mu_1, \mu_2, \mu_3, \mu_4)^\top$ with average values $\sum_i \mu_i/4$ close to the 0.1-0.9 sample quantiles of priors, then we create a set using 0.025 to 0.975 sample quantiles of β_i's selected corresponding to a specific value of plausible charateristics, and the final set is the union of all those sets.

In Table SA-2, when $\gamma = 0$, both techniques produce comparable outcomes. However, when $\gamma = 1$, [CH] sometimes results in a strikingly low coverage rate for $\beta_{\tau,3}$ (for example, when $\tau = 0.2$) while incorporating local plausible characteristics enhances the coverage rate. Table SA-3 explores a different DGP from Table SA-2 with $\theta = (0,1)^\top$ in the DGP and thus $\beta_\tau = 1$; additionally, Table SA-3 also presents findings regarding the 95% confidence sets using the IVQR procedure (see Chernozhukov and Hansen (2005)). We observe similar patterns in Table SA-3 as in Table SA-2.