True and Pseudo-True Parameters

Isaiah Andrews, Harvey Barnhard, Jacob Carlson

April 2, 2024

Abstract

Parameter estimates in misspecified models often converge to pseudo-true parameter values, which minimize a population objective function. Pseudo-true values often differ from quantities of economic interest, raising questions of how, if at all, they are relevant for decision-making. To study this question we consider Bayesian decision-makers facing a population minimum distance problem. Within a class of priors motivated by the minimum distance objective, we characterize prior sequences under which posteriors concentrate on the pseudo-true value. This convergence is fragile to small changes in priors, implying that pseudo-true values are relevant for decision-making only in special cases. Constructive results are nevertheless possible in this setting, and we derive simple confidence intervals that guarantee correct average coverage for the true parameter under every prior in the class we study, with no bound on the magnitude of misspecification.

1 Introduction

Empirical research in economics often begins by positing a model which relates quantities of economic interest to the distribution of observable data. Researchers then use model-implied relationships, together with observed data, to construct estimates or bounds for parameters of interest.

Unfortunately, commonly-used models impose assumptions which are difficult to validate, and which are sometimes rejected outright. For instance, some models impose functional form restrictions such as linearity, or distributional restrictions on latent error terms. Others impose homogeneity across economic agents, or behavioral assumptions such as utility or profit maximization. Finally, methods that aim to uncover causal or structural relationships impose assumptions regarding unconfoundedness of treatment and the scope for spillovers across units. When we have reason to doubt these assumptions, it can be unclear how to interpret model-implied estimates or bounds.

Indeed, absent some restriction on model misspecification, the quantities of interest are necessarily unidentified, and we can learn nothing from data.

An influential literature, including White (1982), Hall and Inoue (2003), Müller (2013), Hansen and Lee (2021), and Andrews and Kwon (2023), studies the problem of inference under model misspecification, and avoids identification problems for the quantity of economic interest θ by shifting the focus to "pseudo-true" parameter values, defined as the minimizers of a population objective function. Under mild conditions these papers show consistency of estimates for the pseudo-true value (in point-identified settings) or identified set (in set-identified ones). Moreover, these papers provide inference results, for example showing asymptotic normality of point estimates and deriving consistent standard errors for the pseudo-true value. These results have been highly influential for empirical practice, with the "sandwich" standard error formula discussed by White (1982), for instance, now widely adopted in the context of maximum likelihood estimation.

While focusing on pseudo-true values allows us to provide statistical guarantees, it leaves open the question of how, if at all, these pseudo-true values relate to the original quantities of economic interest. The literature studying inference on pseudo-true values prominently discusses this tension, with White (1982) writing "[the estimator] converges to a well defined limit, and may or may not be consistent for particular parameters of interest." Similarly, Mueller (2013) writes that "[it] is important to keep in mind that the pseudo-true parameter of the misspecified model must remain the object of interest for ... inference to make sense" and Hansen and Lee (2021) write that "it is difficult to give economic interpretation to pseudo-true parameter values. Consequently, this limits interest in valid inference procedures for pseudo-true values."

This paper revisits the distinction between true and pseudo-true parameter values. To abstract from sampling uncertainty we consider a population minimum-distance problem in which the distribution of the data is perfectly observed. We adopt a decision-theoretic, and specifically Bayesian, perspective to ask under what conditions the posterior distribution for θ, given the distribution of the observable data, concentrates around the pseudo-true parameter. Such concentration implies, under mild conditions on the loss, that Bayes decision rules converge to plug-in rules based on pseudo-true values.

We provide three main results. First, we characterize a class of joint priors for the data distribution and θ such that the posterior density for θ is proportional to a transformation of the minimum distance objective function. This proportionality implies that the minimum distance objective is a sufficient statistic, and hence an optimal way to summarize the data. This is a natural class of priors to consider in the context of minimum distance estimation, since it corresponds to a belief that the

minimum distance objective captures all decision-relevant information. We show that proportionality holds if and only if an implicit prior on the degree of misspecification satisfies a rotation-invariance condition.

Second, we characterize prior sequences in this rotation-invariant class such that the posterior distribution concentrates around the pseudo-true value. These priors assume that the degree of misspecification is negligible, which seems implausible in many economic applications. We further find that concentration is fragile, in the sense that seemingly small changes to the prior lead concentration to fail dramatically. When posterior concentration around the pseudo-true value fails, researchers with different priors on the form and degree of misspecification will have different posteriors for θ, and consequently different Bayes decision rules.

This naturally raises the question of whether it is possible to give positive results under the class of priors we consider. Our third main result constructs confidence intervals that guarantee correct ex-ante coverage of θ under all priors satisfying our rotation-invariance condition. These misspecification-robust confidence intervals have width proportional to the square root of a population J-statistic and so, unlike confidence intervals for pseudo-true values, grow wider as the model fit becomes observably worse, a seemingly natural property for inference procedures in misspecificed models.

The question of inference under model misspecification is closely related to the large literature on inference under partial identification, and the practice of plugging in pseudo-true parameter estimates for decision-making is an instance of what Manski (2021) terms "as-if optimization." In settings where we are concerned with model misspecification, an alternative approach, implemented in various contexts by Conley et al. (2012), Manski and Pepper (2018), Armstrong and Kolesár (2021), and Rambachan and Roth (2023), is to explicitly bound the possible degree of misspecification and derive results which are valid under all data generating processes satisfying this bound, for instance by characterizing the identified set for the quantity of interest. A seemingly natural approach to bounding misspecification leads, however, to the counter-intuitive property that the width of the identified set shrinks as the observable degree of misspecification grows more severe, rather than widening as our confidence intervals do.

The next section introduces our population minimum distance setting and formally defines model misspecification and pseudo-true values. Section 3 introduces the decision problem we study and provides our first main result, characterizing the class of priors such that the posterior for θ depends on the data through the minimum distance objective. Section 4 characterizes sequences of priors in this class such that the posterior concentrates on the pseudo-true value, and shows that this concentration is fragile in important respects. Finally, Section 5 derives our suggested confidence intervals, motivated by an invariance property derived in Section 3, and compares them to identified

sets based on bounds for the magnitude of misspecification.

2 Setting

2.1 Minimum Distance Model

Suppose that for some sample space $\mathcal{D}$ and $\Delta(\mathcal{D})$ the set of distributions on $\mathcal{D}$, a researcher observes a distribution $P \in \mathcal{P} \subseteq \Delta(\mathcal{D})$. This corresponds to the large-sample limit of a setting where the researcher observes a sample of n observations $D_i \in \mathcal{D}$ drawn iid from P, since as $n \to \infty$ they can consistently estimate P from $\{D_i\}_{i=1}^n$. To abstract from sampling uncertainty, we consider the "population problem" where P is directly observed.

Further suppose that the researcher is interested in an economic quantity $\theta \in \mathbb{R}^p$ and that they have a model that implies that the true (P, θ) pair satisfies

$$g(\theta; P) = Y(P) - X(P)\theta = 0 \tag{1}$$

for known functions $Y : \mathcal{P} \to \mathbb{R}^k$ and $X : \mathcal{P} \to \mathbb{R}^{k \times p}$. We assume that $X(P)$ has full column rank, and unless otherwise noted assume that the model is over-identified, with $k > p$. We refer to $g(\theta; P)$ as "moments," though the linear minimum-distance setting we consider here is more general than linear GMM. We focus on linear-in-parameters moments of the form (1) for simplicity, but our exact results for this linear setting will translate to approximate results for models which can be linearly approximated, for instance under local misspecification as studied by Armstrong and Kolesár (2021).

Example: Linear IV As a first example, suppose that $D_i = (Y_i, X_i, Z_i)$ for $Y_i \in \mathbb{R}$ a scalar outcome, $X_i \in \{0, 1\}$ a binary endogenous treatment, and $Z_i \in \mathbb{R}^k$ a vector of k mean-zero exogenous variables, $E[Z_i] = 0$. We assume that these data are generated from a potential outcomes model, where the potential outcomes $Y_i(x, z)$ may in general depend on both X_i and Z_i, and the potential treatments $X_i(z)$ may depend on Z_i. The parameter of interest $\theta \in \mathbb{R}$ is the average treatment effect (ATE),

$$\theta = E[Y_i(1, Z_i) - Y_i(0, Z_i)],$$

which captures the average effect on Y_i from changing X_i from zero to one.

If the researcher wants to estimate a constant-effect linear instrumental variables model with excluded instrument Z_i, this can be justified by assuming that Z_i is excluded from Y_i, $Y_i(x, z) = Y_i(x, z')$ for all (x, z, z'), that the instrument is randomly assigned $Z_i \perp\!\!\!\perp Y_i(\cdot), X_i(\cdot)$, and that treatment effects are constant, $Y_i(1) - Y_i(0) = \theta$ for all i.

Under these assumptions Y_i follows the linear model $Y_i = X_i\theta + \varepsilon_i$, where $\varepsilon_i = Y_i(0)$ and $E[Z_i\varepsilon_i] = 0$. Consequently, θ solves the moment condition (1) for

$$Y(P) = E_P[Z_iY_i], \; X(P) = E_P[Z_iX_i],$$

which are (up to pre-multiplication by $E_P[Z_iZ_i']^{-1}$) equal to the reduced-form and first-stage coefficient vectors in the linear IV model, respectively. $\triangle$

Example: Logit Model As a second example, suppose that $D_i = (Y_i, X_i)$ for $Y_i \in \{0, 1\}$ a binary outcome and $X_i = (1, \tilde{X}_i) \in \mathbb{R}^2$ an exogenous variable, where $\tilde{X}_i \in \{x_1, ..., x_J\}$. If the researcher assumes a logistic regression (i.e. logit) model for Y_i,

$$Y_i = 1\{X_i'\psi > \varepsilon_i\}$$

where $\varepsilon_i \sim \text{Logistic}(0, 1)$ is independent of X_i, then under this model

$$E_P[Y_i|X_i = x] = \Psi(x'\psi)$$

for $\Psi(x) = \frac{e^x}{1+e^x}$ the logistic function or, equivalently,

$$\Psi^{-1}(E_P[Y_i|X_i = x]) = x'\psi$$

for $\Psi^{-1}(x) = \log\left(\frac{x}{1-x}\right)$ the logit function.

We suppose that the object of interest $\theta \in \mathbb{R}^2$ parameterizes the conditional mean of Y given two as-yet-unobserved values of $\tilde{X}_i$,

$$\theta = (\theta_1, \theta_2)' = (\Psi^{-1}(E[Y_i|X_i = (1, x_1^*)]), \Psi^{-1}(E[Y_i|X_i = (1, x_2^*)]))',$$

where $x_1^*, x_2^* \notin \{x_1, ..., x_J\}$. The model implies that θ solves (1) for

$$Y(P) = \begin{pmatrix} \Psi^{-1}(E_P[Y_i|X_i = (1, x_1)]) \\ \vdots \\ \Psi^{-1}(E_P[Y_i|X_i = (1, x_J)]) \end{pmatrix},$$

$$X(P) = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_J \end{pmatrix} = \begin{pmatrix} \frac{x_2^*}{x_2^* - x_1^*} & -\frac{x_1^*}{x_2^* - x_1^*} \\ -\frac{1}{x_2^* - x_1^*} & \frac{1}{x_2^* - x_1^*} \end{pmatrix}. \quad \triangle$$

2.2 Misspecification and Pseudo-True Values

In many contexts researchers are concerned that their models may be misspecified. In the minimum distance setting we consider, this means that at the true (P, θ) pair

$$g(\theta; P) = \eta \neq 0, \tag{2}$$

for η a parameter that describes the impact of misspecification on the moments. As the following examples highlight, we may have $\eta \neq 0$ for a variety of reasons.

Example: Linear IV (Continued) Suppose we maintain the exclusion and independence assumptions for the instruments Z_i, but allow treatment effects to be heterogeneous across units, $\text{Var}(Y_i(1) - Y_i(0)) > 0$. If this treatment effect heterogeneity is correlated with heterogeneity in the first-stage effect $X_i(z) - X_i(z')$, the results of Imbens and Angrist (1994) imply that the linear IV moments are not in general equal to zero at θ. Instead, for β the vector of one-instrument-at-a-time IV estimands (i.e. the IV coefficient using the first instrument by itself, the second by itself, and so on), $\iota \in \mathbb{R}^k$ the vector of ones, and $\circ$ the elementwise product, the implied value of η is

$$\eta = E_P[Z_i Y_i] - E_P[Z_i X_i]\,\theta = (\beta - \theta \cdot \iota) \circ E_P[Z_i X_i] \neq 0.$$

Hence, the model is misspecified in the sense we consider whenever the one-at-a-time IV estimands differ from the average treatment effect. Note that the IV model can thus be misspecified even when we have only a single instrument, $k = 1$: if in this case we further impose the Imbens and Angrist (1994) monotonicity assumption, the IV model will be misspecified if and only if the local average treatment effect (LATE) differs from the ATE.

In this example we focus on misspecification arising from treatment effect heterogeneity, but our framework is sufficiently general to accommodate many other ways in which the researcher's assumptions could fail. For instance, if the exclusion restriction fails, so $Y_i(x, z) \neq Y_i(x, z')$ for some (x, z, z'), or independence fails and $Z_i \not\perp\!\!\!\perp (Y_i(\cdot), X_i(\cdot))$, each of these will will imply a particular form for η. $\triangle$

Example: Logit Model (Continued) The logit model may be misspecified for a variety of reasons, for instance because the linear threshold model is incorrect and x in fact enters nonlinearly, $Y_i = 1\{h(X_i) > \varepsilon_i\}$, or because the linear threshold model is correct but the residual ε_i does not follow a logistic distribution. Whatever the reason

for misspecification, we will have

$$\eta = \begin{pmatrix} \Psi^{-1}\left(E_P\left[Y_i|X_i=(1,x_1)\right]\right) \\ \vdots \\ \Psi^{-1}\left(E_P\left[Y_i|X_i=(1,x_J)\right]\right) \end{pmatrix} - \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_J \end{pmatrix} \begin{pmatrix} \frac{x_2^*}{x_2^*-x_1^*} & -\frac{x_1^*}{x_2^*-x_1^*} \\ -\frac{1}{x_2^*-x_1^*} & \frac{1}{x_2^*-x_1^*} \end{pmatrix} \theta \neq 0,$$

where the conditional expectations $E_P[Y_i|X_i=(1,x_j)]$ will depend of the precise form of misspecification. $\triangle$

If we allow $\eta \neq 0$ and impose no other restrictions, identification of θ is hopeless, since any value of θ is compatible with any distribution P. To avoid such a pessimistic conclusion one route pursued in the literature, including in Conley et al. (2012), Manski and Pepper (2018), Masten and Poirier (2020), Armstrong and Kolesár (2021), and Rambachan and Roth (2023), is to consider bounded relaxations of the model.

First, note that for a P-dependent positive-definite weighting matrix $W(P)$, the model is correctly specified if and only if the W-weighted norm of the misspecification parameter η is equal to zero

$$\|\eta\|_W := \sqrt{\eta' W(P)\,\eta} = 0.$$

To allow the possibility of misspecification, the researcher could thus relax this assumption, and assume only that $\|\eta\|_W$ is bounded above by some known constant d. The identified set for θ under this relaxation is then the set of values θ such that the minimum distance objective function

$$Q_W(\theta;P) := \|g(\theta;P)\|_W^2$$

takes a value smaller than d^2

$$\Theta_I(P,d) := \left\{\theta : Q_W(\theta;P) \leq d^2\right\}. \tag{3}$$

While this approach requires the researcher to specify the norm bound d, the data do contain some information about this quantity. Specifically, when $d^2 < J_W(P)$ for $J_W(P) = \min_\theta Q_W(\theta;P)$ the population analog of the J-statistic (Hansen, 1982), $\Theta_I(P,d)$ is empty, so the data reject the assumption that $\|\eta\|_W < d$. Thus, the data imply lower, but not in general upper, bounds on the degree of misspecification.

Another common practice when we are concerned with model misspecification is to focus on pseudo-true parameter values (c.f. White 1982, Müller 2013, Hansen and Lee 2021, and Andrews and Kwon 2023), which are defined as the minimizers of a population objective function. In our setting, the pseudo-true parameter corresponds

to the value of θ at which $J_W(P)$ is attained, and takes a simple form

$$\theta_W(P) = \arg\min_\theta Q_W(\theta; P) = (X(P)'W(P)X(P))^{-1}X(P)'W(P)Y(P).$$

Note that $\theta_W(P)$ is equal to the coefficient from a generalized least squares regression of $Y(P)$ on $X(P)$, weighting by $W(P)$.

Example: Linear IV (Continued) In the linear IV model with treatment effect heterogeneity it is common to focus attention on the two-stage least squares (TSLS) estimand, which corresponds to the pseudo-true value using the TSLS weighting matrix $W(P) = E_P[Z_i Z_i']^{-1}$, and can be interpreted as a LATE under appropriate assumptions (Angrist and Imbens, 1995). $\triangle$

Example: Logit Model (Continued) In the logit model with misspecification, we cannot choose θ to match the full set of observed conditional means $E_P[Y_i|X_i = x]$. The weighting matrix governs how we prioritize matching different elements of this vector, and one natural choice is to take $W(P)$ to be the diagonal matrix with jth diagonal element equal to the probability that $\tilde{X}_i = x_j$, $E_P[1\{\tilde{X}_i = x_j\}]$, which prioritizes matching the conditional mean for X_i values which are more common in the population.[1] $\triangle$

The pseudo-true parameter corresponds exactly to the identified set with $d^2 = J_W(P)$, $\Theta_I(P, J_W(P)) = \{\theta_W(P)\}$. Hence, if a researcher assumes the true parameter value is equal to the pseudo-true, this is the same as assuming that the degree of misspecification, measured in the norm $\|\cdot\|_W$ is as small as it can possibly be given the observed distribution P. If they instead allow the possibility that θ and $\theta_W(P)$ are different, then as discussed in the introduction it is not obvious how, if at all, the pseudo-true value $\theta_W(P)$ relates to the economic questions that motivate the analysis in the first place. Consequently, it is unclear when we would want to estimate pseudo-true values. The following two sections consider this question from a decision-theoretic perspective, providing conditions under which optimal decisions depend on the data through (i) the population minimum distance objective $Q(\theta; P)$ and (ii) the pseudo-true value $\theta_W(P)$ in particular.

[1]Interestingly, this can be shown to correspond to the limit of the optimal minimum distance weighting matrix for this model under correct specification.

3 Optimal Decisions and Minimum Distance

Researchers are often interested in estimating economic parameters in order to inform decisions by policymakers, businesses, or households. It is not obvious that pseudo-true values are suitable for this purpose. To explore this question, we adopt a decision-theoretic perspective and ask under what conditions Bayesian decision-makers would be willing to base their decisions on minimum distance methods and the pseudo-true values they generate.

3.1 Decision Problem

Consider a decision-maker who has to choose an action a from a set of possible actions $\mathcal{A}$. After choosing action a, the decision-maker suffers a loss $L(a, \theta)$ that depends on the action taken and the true value for θ. If θ were known the optimal action for the decision-maker would be to simply choose $a \in \arg\min_{a \in \mathcal{A}} L(a, \theta)$.

In practice θ is unknown, and the decision-maker instead observes only the distribution $P \in \mathcal{P}$ of the data. Hence, the decision-maker must select a decision rule $\delta : \mathcal{P} \to \mathcal{A}$ that maps data distributions into actions. The decision-maker prefers decision rules δ that yield a lower loss, $L(\delta(P), \theta)$, but when θ cannot be uniquely determined based on P (e.g. when θ is set-identified due to model misspecification), different decision rules δ will perform best at different (P, θ) pairs, and there generally will not be a uniformly best choice.

To select among possible decision rules in settings without a uniformly best rule, the decision-maker necessarily trades off performance across different (P, θ) pairs. One way to formalize such tradeoffs is to consider Bayes decision rules, which weight losses across different (P, θ) pairs according to a prior $\pi \in \Delta(\mathcal{P} \times \mathbb{R}^p)$. The Bayes decision rule δ_π minimizes the average loss under the prior,

$$\delta_\pi \in \arg\min_{\delta} \int L(\delta(P), \theta) d\pi(P, \theta).$$

To compute δ_π, it suffices to minimize the posterior expected loss at each P,

$$\delta_\pi(P) \in \arg\min_{a \in \mathcal{A}} \int L(a, \theta) \pi(\theta | P) d\theta,$$

where for simplicity we assume the posterior for $\theta|P$ is continuous and write $\pi(\theta|P)$ for the posterior density.

Example: Linear IV (Continued) As in Andrews and Shapiro (2021), suppose the decision-maker needs to set a tax or subsidy $a \in \mathbb{R}$ for the treatment, where $a > 0$ denotes a subsidy, while $a < 0$ denotes a tax, and that the loss is $L(a, \theta) = (a - \theta)^2$.

That is, the optimal subsidy level is equal to the average treatment effect, while the loss increases quadratically as the subsidy departs from the ATE. $\triangle$

3.2 Minimum-Distance Priors

The posterior density $\pi(\theta|P)$ summarizes all decision-relevant information about the parameter θ given the data. To connect minimum distance methods and Bayes decision rules, we consider a class of decision-makers for whom the minimum distance objective function is a sufficient statistic, in the specific sense that $\pi(\theta|P)$ is proportional to a function of $(Q_W(\theta;P), W(P), X(P))$. For such priors the minimum distance objective contains all decision-relevant information, so and is the natural basis for decisionmaking.

Assumption 1 *The conditional prior $\pi(Y(P), \theta|X(P), W(P))$ is absolutely continuous for all $X(P), W(P)$. Moreover, for all $P \in \mathcal{P}$,*

$$\pi(\theta|P) \propto h(Q_W(\theta;P), W(P), X(P), \theta)$$

for a non-negative function h.

The first part of the assumption is a continuity requirement that is imposed primarily for convenience and could be weakened. The second part of the assumption connects the posterior distribution to the minimum distance objective, and is weaker than assuming that $\pi(\theta|P)$ is proportional to a function $h(Q_W(\theta;P), \theta)$ as in the Gibbs posterior distributions studied in the statistics and machine learning literature (e.g. Catoni 2007, Alquier et al. 2016, Bissiri et al. 2016, Martin and Syring 2022) and the quasi-Bayesian approach of Chernozhukov and Hong (2003). Under Assumption 1, providing the decision-maker with $(Q_W(\cdot|P), W(P), X(P))$ is as good as providing them with the full data. By contrast, when this sufficiency fails to hold minimum-distance methods sacrifice decision-relevant information and so may not be appropriate. Hence, we view Assumption 1 as a reasonable restriction in settings where researchers are considering minimum distance methods.

Assumption 1 immediately implies that the posterior density $\pi(\theta|P)$ depends on the data only through $(W(P), X(P), Y(P))$.

Lemma 1 *Under Assumption 1, $\pi(\theta|P) = \pi(\theta|W(P), X(P), Y(P))$.*

Example: Linear IV (Continued) Focusing on the case where $W(P)$ is the two-stage least squares weighting matrix $W(P) = E_P[Z_i Z_i']^{-1}$, Lemma 1 implies that the decision-maker's posterior for the ATE depends on the data only through the reduced-form and first stage regression coefficients, together with the covariance matrix of the instruments. This rules out, for instance, priors such that the decision-maker's beliefs

about the ATE are informed by higher moments of P. $\triangle$

Assumption 1 further implies that the posterior density $\pi(\theta|P)$ can be expressed in terms of the minimum distance moments (1). In particular, the first part of Assumption 1 implies that the conditional priors for θ given $(X(P), W(P))$ and η given $(W(P), X(P), \theta)$ are continuous. For brevity, let $\pi_\theta(\theta) := \pi(\theta|W(P), X(P))$ and $\pi_\eta(\eta|\theta) := \pi(\eta|W(P), X(P), \theta)$ denote their densities with respect to Lebesgue measure. With this notation the posterior density of θ given P is

$$\pi(\theta|P) = \frac{\pi_\theta(\theta)\pi_\eta(Y(P) - X(P)\theta|\theta)}{\int \pi_\theta(\theta)\pi_\eta(Y(P) - X(P)\theta|\theta)d\theta} = \frac{\pi_\theta(\theta)\pi_\eta(g(\theta; P)|\theta)}{\int \pi_\theta(\theta)\pi_\eta(g(\theta; P)|\theta)d\theta}. \tag{4}$$

Examining this expression, we see that it resembles the posterior in a finite-sample problem with parameter θ, prior π_θ, and likelihood π_η. Consistent with this resemblance the posterior $\pi(\theta|P)$ will be non-degenerate with non-trivial uncertainty about the true value of θ even though the data distribution P in our problem is perfectly known. This reflects the fact that θ is not point-identified, so since the conditional prior π_η on the degree of misspecification is non-dogmatic the decision-maker remains uncertain about θ even after observing P.

Assumption 1 also restricts the form of $\pi_\eta(\eta|\theta)$, which describes the prior distribution for the moments evaluated at the true parameter value θ. In particular, Assumption 1 implies that the prior density at η conditional on $(W(P), X(P), \theta)$ depends only on $\eta' W(P)\eta$, $\pi_\eta(\eta|\theta) \propto f(\eta' W(P)\eta|\theta)$, where the function f may also vary with $(W(P), X(P))$.

Lemma 2 *Assumption 1 implies that*

$$\pi_\eta(\eta|\theta) \propto f(\eta' W(P)\eta|\theta)$$

for a non-negative function $f(u|\theta) := f(u|W(P), X(P), \theta)$ *with* $\int f(\eta'\eta|\theta)d\eta < \infty$ *for all* θ, P.

Lemma 2 implies that the prior density $\pi_\eta(\eta|\theta)$ is invariant to rotation of $W(P)^{\frac{1}{2}}\eta$, in the sense that for any η, $\tilde{\eta}$ such that $W(P)^{\frac{1}{2}}\eta = OW(P)^{\frac{1}{2}}\tilde{\eta}$ for a rotation matrix O, the prior density is the same at η and $\tilde{\eta}$, $\pi_\eta(\eta|\theta) = \pi_\eta(\tilde{\eta}|\theta)$. The density $\pi_\eta(\eta|\theta)$ is thus constant on the ellipsoids $\{\eta : \eta' W'(P)\eta = C\}$ for all constants C, from which it follows that $\pi_\eta(\eta|\theta)$ is an elliptically-contoured distribution (Muirhead, 1982).

Example: Linear IV (Continued) Recall that $\eta = (\beta - \theta \cdot \iota) \circ E_P[Z_i X_i]$ measures the difference between the average treatment effect and the vector of one-instrument-at-a-time IV estimands. One example of an elliptically-contoured distribution in this

setting takes $\eta|\theta, W(P) \sim N(0, W(P)^{-1})$, which corresponds to $f(u|\theta) = \exp(-\frac{1}{2}u)$. Under this prior, each LATE is equal to the ATE plus a mean-zero noise term, where the covariance matrix of the noise is determined by $W(P)$ and the first stage $E_P[Z_i X_i]$. There are many other rotation-invariant priors, however, including a multivariate t prior with ν degrees of freedom, where $f(u|\theta) = \left(1 + \frac{1}{\nu}u\right)^{-\frac{\nu+k}{2}}$, and an elliptically contoured power law which takes $f(u|\theta) = u^{-\kappa-1}$ for $\kappa > k-1$. $\triangle$

Together, the conclusions of Lemmas 1 and 2 imply that the posterior density $\pi(\theta|P)$ is proportional to a function of $(Q_W(\theta; P), W(P), X(P))$, as required by Assumption 1. The next proposition summarizes this line of reasoning.

Proposition 1 *Assumption 1 holds if and only if for all $P \in \mathcal{P}$,*

$$\pi(\theta|P) = \frac{\pi_\theta(\theta) f(Q_W(\theta; P)|\theta)}{\int \pi_\theta(\theta) f(Q_W(\theta; P)|\theta) d\theta},$$

for a non-negative function $f(u|\theta) := f(u|W(P), X(P), \theta)$ with $\int f(\eta'\eta|\theta)d\eta < \infty$ for all θ, P.

As discussed above, the posterior distribution will typically feature non-trivial uncertainty about the parameter θ.

4 Concentration-Inducing Priors

We next show that for particular sequences of priors satisfying Assumption 1, the corresponding posterior distributions concentrate around the pseudo-true parameter value $\theta_W(P)$. This concentration in turn implies that Bayes decision rules converge to plug-in decision rules for a large class of loss functions. The prior sequences we consider imply that the model is misspecified with probability one (in the sense that $\|\eta\|_W > 0$ almost surely under the prior), but take the expected *magnitude* of misspecification to zero. Hence, while these priors allow the possibility of misspecification, they assume that the degree of misspecification is arbitrarily small.

An assumption that the degree of misspecification is arbitrarily small is unreasonable in many economic applications. Moreover, we show that concentration around the pseudo-true value is fragile. First, we show that if we take our concentration-inducing prior sequences and mix them, to an arbitrarily small degree, with any fixed full-support prior on θ and the degree of misspecification η, concentration around the pseudo-true value immediately fails whenever $J_W(P) > 0$. Second, we show that even under prior sequences which imply a vanishing degree of misspecification, concentration around the pseudo-true value requires that the prior on η be sufficiently thin-tailed.

4.1 Posterior Concentration

To provide sufficient conditions for posterior concentration around the pseudo-true parameter value we first assume that the prior density $\pi_\eta(\eta|\theta)$ is independent of θ and thin-tailed, in the sense that $f(u|\theta) = f(u)$ decays at a faster-than polynomial rate as $u \to \infty$.

Assumption 2 *$f(u|\theta) = f(u)$, where $f(u)$ is strictly positive, continuous and non-increasing in u, and satisfies $\lim_{u\to\infty} \frac{f(au)}{f(u)} = 0$ for all $a > 1$.*

This assumption is important for the concentration results derived in this section and holds, for instance, when $\pi(\eta|\theta)$ is a normal density.

Example: Linear IV (Continued) If $\pi_\eta(\eta|\theta)$ is a $N(0, W(P)^{-1})$ density, then since $\frac{\exp\left(-\frac{a}{2}u\right)}{\exp\left(-\frac{1}{2}u\right)} = \exp\left(\frac{1-a}{2}u\right)$ and $a > 1$, Assumption 2 holds. $\triangle$

Under sequences of priors that satisfy Assumptions 1 and 2 and take the degree of misspecification to be small, the posterior distribution concentrates on the pseudo-true value. To show this formally, we consider a scale family of priors on the misspecification parameter η, $\pi_{\eta,c}(\eta|\theta) \propto f(\frac{1}{c}\eta' W(P)\eta)$. The scale parameter c controls the magnitude of misspecification implied by the prior, and the prior variance of η is proportional to c. Our main result in this section considers the behavior of the posterior as the scale parameter becomes small, $c \to 0$, corresponding to priors that assume a vanishing degree of misspecification.

Proposition 2 *Suppose Assumptions 1 and 2 hold. For any continuous $\pi_\theta(\theta)$ with $\pi_\theta(\theta_W(P)) > 0$, the posterior*

$$\pi_c(\theta|P) = \frac{\pi_\theta(\theta) f(\frac{1}{c} Q_W(\theta; P))}{\int \pi_\theta(\theta) f(\frac{1}{c} Q_W(\theta; P)) d\theta}$$

concentrates on $\theta_W(P)$ as $c \to 0$. For $B_\varepsilon(\theta_W(P)) = \{\theta : \|\theta_W(P) - \theta\| < \varepsilon\}$,

$$\lim_{c\to 0} \int 1\{\theta \notin B_\varepsilon(\theta_W(P))\} d\pi_c(\theta|P) = 0 \text{ for all } \varepsilon > 0.$$

Proposition 2 shows that for priors satisfying Assumptions 1 and 2 where the degree of misspecification is small, the posterior distribution concentrates on the pseudo-true parameter value $\theta_W(P)$. This is entirely expected when $J_W(P) = 0$, since in this case the data provide no evidence of misspecification and priors with $c \to 0$ put vanishing probability on substantial misspecification. When $J_W(P) > 0$, by contrast, the data imply non-trivial misspecification but Proposition 2 shows that the posterior continues to concentrate.

Concentration of the posterior $\pi_c(\theta|P)$ translates to convergence of the Bayes decision rule

$$\delta_{\pi_c}(P) \in \arg\min_{a\in\mathcal{A}} \int L(a,\theta) d\pi_c(\theta|P),$$

under conditions on the decision problem.

Proposition 3 *Suppose that $\mathcal{A}$ is compact under some metric d, that $\sup_{a,\theta} L(a,\theta) < \infty$, that $\sup_{a,a',\theta} |L(a,\theta) - L(a',\theta)| < \lambda \cdot d(a,a')$ for some $\lambda > 0$, and that the loss $L(a,\theta)$ has a unique minimum for all θ. Then as $c \to 0$,*

$$\delta_{\pi_c}(P) \to \arg\min_{a\in\mathcal{A}} L(a, \theta_W(P)).$$

Proposition 3 shows that for bounded loss functions that are Lipschitz in a, Bayes decision rules corresponding to the priors we study converge to plug-in decision-rules based on the pseudo-true parameter value. This result is useful for a number of reasons. First, it shows that plug-in decision rules using the pseudo-true parameter value correspond to the limit of a sequence of Bayes decision rules for a large class of loss functions, providing one justification for such plug-in rules. Second, it shows that the pseudo-true parameter value $\theta_W(P)$ is a sufficient statistic for communication with an audience whose priors take the limiting form we consider: a researcher looking to summarize the data for such an audience is justified in reporting only the pseudo-true parameter value, since it allows audience members to compute the optimal decision for whatever loss function they have, provided that loss satisfies the conditions of Proposition 3.

The conditions on the loss function in Proposition 3 are somewhat restrictive and rule out squared error loss on an unbounded domain. These conditions only are sufficient and not necessary for convergence of decision rules, however, and we can obtain convergence in many settings with unbounded loss functions by using additional structure for the loss function and prior.

Example: Linear IV (Continued) Proposition 3 does not apply in this example, because the loss $L(a,\theta)$ is unbounded and, moreover, is not Lipschitz in a. Nonetheless, if $\pi_\eta(\eta|\theta)$ corresponds to a $N\left(0, \frac{1}{c}W(P)^{-1}\right)$ distribution while the prior on θ is flat, the posterior density is

$$\begin{aligned}\pi_c(\theta|P) &= N\left(\theta_W(P), c\cdot(E_P[Z_iX_i]'W(P)E_P[Z_iX_i])^{-1}\right)\\ &= N\left(\theta_W(P), c\cdot(X(P)'W(P)X(P))^{-1}\right).\end{aligned}$$

Hence, for all c the posterior distribution is a normal centered at $\theta_W(P)$ with variance proportional to c. Consistent with Proposition 2, this posterior converges weakly to a point mass at $\theta_W(P)$ as $c \to 0$. Moreover, despite the conditions of Proposition 3 not

holding in this example, the Bayes decision rule $\delta_{\pi_c}(P)$ is equal to $\theta_W(P)$ for all c, so the Bayes decision rule always agrees with the plug-in rule based on the pseudo-true value. $\triangle$

4.2 Posterior Concentration is Fragile

This section shows that the convergence established by Propositions 2 and 3 is fragile in two important respects.

Fragility to Prior Contamination First, we show that if we mix the concentration-inducing priors studied in the previous section with any fixed, full support prior for $\eta|\theta$, posterior concentration fails when $J_W(P) > 0$.

Proposition 4 *Consider conditional priors of the form*

$$\pi^{\phi}_{\eta,c}(\eta|\theta) = (1-\phi)\pi_{\eta,c}(\eta|\theta) + \phi\pi^*_{\eta}(\eta|\theta)$$

for any full-support conditional prior $\pi^*_\eta(\eta|\theta)$ and $\phi \in (0,1)$. If $J_W(P) > 0$, then under Assumption 2, for any $\pi_\theta(\theta)$ the resulting posterior satisfies

$$\lim_{c\to 0} \pi^{\phi}_c(\theta|P) = \frac{\pi_\theta(\theta)\pi^*(Y(P) - X(P)\theta|\theta)}{\int \pi_\theta(\theta)\pi^*(Y(P) - X(P)\theta|\theta)d\theta}.$$

If instead $J_W(P) = 0$, $\pi_\theta(\theta)$ is continuous, and $\pi_\theta(\theta_W(P)) > 0$, then

$$\lim_{c\to 0} \int 1\{\theta \notin B_\phi(\theta_W(P))\} d\pi^{\phi}_c(\theta|P) = h(\phi),$$

where $\lim_{\phi\to 0} h(\phi) = 0$.

Proposition 4 shows that if $J_W(P) > 0$ and we contaminate the concentration inducing prior $\pi_{\eta,c}(\eta|\theta)$, to an arbitrarily small extent, with any full-support prior π^* for $\eta|\theta$ then the posterior converges to a the same limit as if we had set $\pi_\eta(\eta|\theta) = \pi^*_\eta(\eta|\theta)$. By contrast, when $J_W(P) = 0$, the $c \to 0$ limiting posterior continues to have a point-mass at $\theta_W(P)$, where the mass assigned to this point converges to one when $\phi \to 0$. Hence, in the $\phi \to 0$ limit we obtain the same concentration result as in Proposition 2.

Proposition 4 can be interpreted in terms of pre-testing for model specification: in the case where $J_W(P) > 0$, the data imply that the model is non-trivially misspecified. By contrast, while the priors $\pi_{\eta,c}(\eta|\theta)$ imply that the model is misspecified with probability one, as $c \to 0$ they imply that the degree of misspecification is arbitrarily small.

When we allow the possibility that η is instead drawn from a fixed full-support distribution π^*_η, for c sufficiently small the data provide arbitrarily strong support for

$\pi^*_\eta(\eta|\theta)$ over $\pi_{\eta,c}(\eta|\theta)$. Loosely speaking, the concentration-inducing prior $\pi_{\eta,c}(\eta|\theta)$ is rejected in favor of the full-support prior $\pi^*_\eta(\eta|\theta)$. By contrast, when $J_W(P) = 0$ the data are consistent with correct specification of the model. Moreover, as $c \to 0$ the density $\pi_{\eta,c}(0|\theta)$ diverges to infinity while $\pi^*_\eta(0|\theta)$ is fixed. Consequently, the posterior $\pi^\phi_c(\theta|P)$ for $\phi > 0$ concentrates some of its mass around $\theta_W(P)$ as $c \to 0$, where the exact amount of mass is controlled by ϕ.[2]

Fragility to Heavy Tails Second, we show that even under prior sequences such that the degree of misspecification goes to zero, the concentration obtained in Proposition 2 relies on thin tails for $\pi_\eta(\eta|\theta)$. To illustrate this point, we show that posterior concentration around the pseudo-true value fails in two examples with heavy-tailed priors.

Example: Posterior Non-Concentration with t Prior Suppose that our prior on η corresponds to a multivariate t distribution centered at zero with scale matrix $W(P)^{-1}$ and $\tilde{\nu}$ degrees of freedom, $f(u) \propto \left(1 + \frac{1}{\tilde{\nu}}u\right)^{-\frac{\tilde{\nu}+k}{2}}$. Provided $J_W(P) > 0$, if we define $\nu = \tilde{\nu} + k - p$ and

$$\Sigma(P) = J_W(P)\left(\nu X(P)'W(P)X(P)\right)^{-1}$$

it follows that

$$\lim_{c\to 0} \pi_c(\theta|P) \propto \pi_\theta(\theta)\left(1 + \nu^{-1}(\theta - \theta_W(P))'\Sigma(P)^{-1}(\theta - \theta_W(P))\right)^{-(\nu+p)/2},$$

where the second term is the density for a multivariate t distribution centered at $\theta_W(P)$, with scale matrix $\Sigma(P)$ and ν degrees of freedom. Consequently, the $c \to 0$ limiting posterior corresponds to updating the prior π_θ based on observing $\theta_W(P) \sim t_\nu(\theta, \Sigma(P))$.

Note that the degrees of freedom in the "likelihood," ν, is equal to the degrees of freedom in the misspecification prior π_η plus the degree of over-identification, so a higher degree of over-identification leads to thinner tails for the posterior all else equal. The scale parameter in the "likelihood," $\Sigma = J_W(P)\left(\nu X(P)'W(P)X(P)\right)^{-1}$, is increasing in the J-statistic so cases where the moment conditions are observed to be more badly violated lead to a more uncertain posterior, all else equal. $\triangle$

The tail thickness of $f(\cdot)$ matters because it determines beliefs about the total level of misspecification conditional on a given value of the J-statistic. To see this, let us again assume that $\eta|\theta \sim \pi_{\eta,c}(\eta|\theta)$ and consider the conditional distribution of $\|\eta\|_W$ conditional on the norm of η exceeding a threshold τ, $\|\eta\|_W|\left(\tau < \|\eta\|_W\right)$. For thin-

[2]We thank Jesse Shapiro for pointing out this connection.

tailed priors (i.e. those satisfying Assumption 2), one can show that for all $a > 1$ and all positive constants $\tau > 0$,

$$\lim_{c \to \infty} Pr_{\pi_{\eta,c}} \left\{ \|\eta\|_W \geq a \cdot \tau \middle| \left(\tau \leq \|\eta\|_W \right) \right\} \to 0. \tag{5}$$

Hence, even in the case where the model is known to be misspecified, as $c \to 0$ thin-tailed priors imply that the degree of *additional* misspecification, beyond that implied by the lower bound, is negligible. Recall, however, that the J-statistic is itself a lower bound on the total degree of misspecification, $\|\eta\|_W \geq J_W(P)$. Consistent with this, when $c \to 0$ thin-tailed priors imply that the total degree of misspecification must not be much larger than that suggested by the J-statistic, and thus that θ must the close to $\theta_W(P)$, which is the unique parameter value compatible with $\|\eta\|_W = J_W(P)$.

By contrast, for f corresponding to a multivariate t distribution we have that

$$\lim_{c \to \infty} Pr_{\pi_{\eta,c}} \left\{ \|\eta\|_W \geq a \cdot \tau \middle| \left(\tau \leq \|\eta\|_W \right) \right\} \to p(a). \tag{6}$$

for a fixed, nonzero function $p(\cdot)$ that does not depend on the misspecification lower bound τ. Consequently, in this case the researcher's belief about the total degree of misspecification is non-degenerate and, once c is sufficiently small, scales proportionally with τ.

While t-distributed priors for $\eta|\theta$ correspond to updates via a t likelihood in the $c \to 0$ limit, if we instead consider multivariate power law prior then dependence on c vanishes entirely.

Example: Posterior Non-Concentration with Power Law Prior Suppose $f(x) = x^{-\alpha}$ for $\alpha > k$. Then for $\nu = 2\alpha - p$,

$$\Sigma(P) = J_W(P) \left(\nu X(P)' W(P) X(P) \right)^{-1},$$

and all c,

$$\pi_c(\theta|P) \propto \pi(\theta) \left(1 + \nu^{-1} (\theta - \theta_W(P))' \Sigma^{-1} (\theta - \theta_W(P)) \right)^{-(\nu+p)/2}$$

which corresponds to the posterior distribution from observing $\theta_W(P) \sim t_\nu(\theta, \Sigma(P))$. We again see that a higher degree of over-identification leads to thinner tails for the posterior, while a larger J-statistic leads to a more dispersed posterior. $\triangle$

5 Confidence Sets Based on Rotation-Invariance

Outside of the restrictive cases where $\pi(\theta|P)$ concentrates on the pseudo-true value, researchers will have non-trivial uncertainty about the true value of θ even after observing P, where the exact posterior will depend on the details of the prior. Despite this dependence of posterior beliefs on the prior, in this section we construct confidence intervals that have correct coverage ex-ante for the true value θ under all priors satisfying Assumption 1. Since as we previously argued this class has a close connection to minimum distance methods, we view these confidence sets as a natural summary for misspecification-driven uncertainty in settings where researchers adopt a minimum distance approach.

To state this result, suppose the researcher is interested in inference on a linear combination of the elements of θ, $v'\theta$ for $v \in \mathbb{R}^p$. For $t^*_{k-p,1-\frac{\beta}{2}}$ the level $1-\frac{\beta}{2}$ critical value for a standard t distribution with $k-p$ degrees of freedom, $\theta_W(P) = \arg\min_\theta Q_W(\theta; P)$ the pseudo-true value, $J_W(P) = \min_\theta Q_W(\theta; P)$ the population J-statistic, and $H(P) = X(P)'W(P)X(P) = \frac{\partial^2}{\partial\theta\partial\theta'}Q_W(\theta; P)$ the Hessian of the minimum distance objective, define the confidence interval

$$CI(P) := \left[v'\theta_W(P) \pm \sqrt{\frac{J_W(P)}{k-p}} \cdot \sqrt{v'H(P)^{-1}v} \cdot t^*_{k-p,1-\frac{\beta}{2}} \right]. \tag{7}$$

This confidence interval has correct coverage conditional on $(W(P), X(P))$, and thus correct ex-ante coverage, for all priors satisfying Assumption 1.

Proposition 5 *For any prior π such that Assumption 1 holds,*

$$Pr_\pi\left\{v'\theta \in CI(P)|W(P), X(P)\right\} = Pr_\pi\left\{v'\theta \in CI(P)\right\} = 1-\beta.$$

The confidence interval (7) has a number of interesting features. While it is centered at the pseudo-true value, its width is governed by (i) the Hessian of the minimum distance objective function $H(P)$ and (ii) the population J-statistic $J_W(P)$. The fact that a smaller Hessian leads to wider confidence intervals resembles many other inference problems, though ours is unusual in that our intervals are derived in a population problem and reflect uncertainty due to misspecification, rather than sampling uncertainty. The dependence on the population J-statistic is non-standard, but seems intuitively reasonable given the connection to model misspecification.

It is important to note that the notion of coverage considered in Proposition 5 is non-standard. Since we consider the population problem where P is observed, the frequentist coverage is either zero or one, $Pr_{(P,\theta)}\{v'\theta \in CI(P)\} \in \{0, 1\}$. In Proposition

5, we instead consider average coverage under π,

$$Pr_\pi\left\{v'\theta \in CI(P)\right\} = \int Pr_{(P,\theta)}\left\{v'\theta \in CI(P)\right\} d\pi(P,\theta),$$

which measures ex-ante coverage probability under the prior. Proposition 5 thus establishes that any Bayesian whose prior satisfies Assumption 1 thinks, prior to seeing the data, that the interval $CI(P)$ has coverage $1-\beta$. In this sense, (7) is a confidence interval with correct average coverage under the class of priors satisfying Assumption 1.

We provide intuition for Proposition 5 from two perspectives, first establishing a connection to generalized least squares and then showing that this confidence interval corresponds to a credible set under an improper power law prior.

Regression Interpretation Note that for

$$(\tilde{Y}, \tilde{X}, \tilde{\eta}) = W(P)^{\frac{1}{2}}(Y(P), X(P), \eta)$$

we can write the minimum distance model allowing for misspecification, as

$$\tilde{Y}(P) = \tilde{X}(P)\theta + \tilde{\eta}, \tag{8}$$

where under all priors π satisfying Assumption 1, the conditional distribution of $\tilde{\eta}|\tilde{X}(P), \theta$ is rotation-invariant, $\tilde{\eta}|\tilde{X}(P), \theta \sim O\tilde{\eta}|\tilde{X}(P), \theta$ for all orthonormal matrices O. Hence, for any prior satisfying Assumption 1 the problem of inference on θ conditional on $\tilde{X}(P)$ reduces to that of inference on the coefficient in a regression with a rotation-invariant error. Since the validity of t-statistics relies only on rotation-invariance of the error and not the exact distribution, t-tests provide valid tests for scalar coefficients $c'\theta$ in this setting. The confidence interval (7), however, is exactly the t-statistic confidence interval derived from (8). If we instead want a confidence interval for a multi-dimensional combination of the coefficients θ, the analogous approach based on F-statistics is also valid.

Bayesian Interpretation Recall that under the multivariate power-law prior for $\eta|\theta$, $f(x) = x^{-\alpha}$, the posterior $\pi_c(\theta|P)$ corresponds to a t distribution with $2\alpha - p$ degrees of freedom. Consequently, the confidence interval (7) corresponds to a posterior credible set under a flat prior on θ and a multivariate power law prior on η with $\alpha = k/2$. This is an improper prior for η, since in this case $\int \pi_\eta(\eta|\theta)d\eta = \int f(\eta' W(P)\eta|\theta)d\eta = \infty$, but yields a proper posterior $\pi(\theta|P)$. Viewed from this perspective, Proposition 7 shows that there exists an improper conditional prior on $\theta, \eta|W(P), X(P)$ whose credible sets have correct average coverage under all priors satisfying Assumption 1.

5.1 Comparison to Norm-Bound Identified Sets

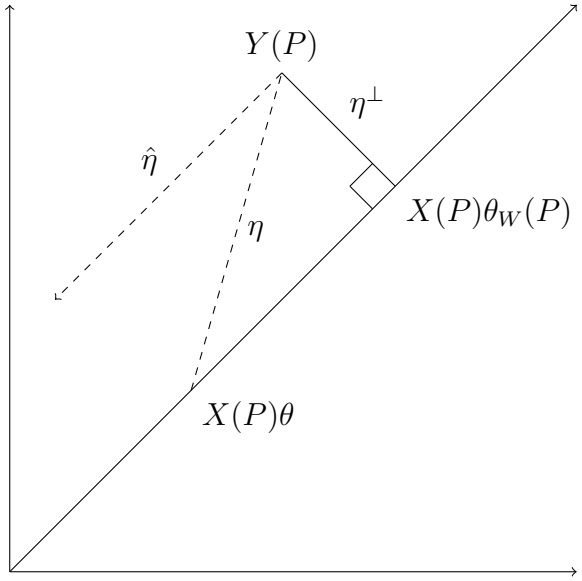

Figure 1: Relationship between the pseudo-true value and true θ in an example with scalar θ. The solid diagonal line represents the column space of $X(P)$. The vector $\eta^\perp$ perpendicular to the column space captures the detectable component of the misspecification vector, while vector $\hat{\eta}$ parallel to the column space captures the undetectable component of misspecification.

We next compare the behavior of the confidence set (7) to the identified set (3) constructed under the assumption that $\|\eta\|_W \leq d$. To facilitate this comparison, note that we can decompose

$$\tilde{\eta} := W(P)^{\frac{1}{2}}\eta = M(P)\tilde{\eta} + (I - M(P))\tilde{\eta} := \hat{\eta} + \eta^\perp$$

for

$$M(P) = \tilde{X}(P)(\tilde{X}(P)'\tilde{X}(P))^{-1}\tilde{X}(P)'$$

the projection matrix onto $\tilde{X}(P) = W(P)^{\frac{1}{2}}X(P)$. Here $\hat{\eta}$ and $\eta^\perp$ are the projection of $\tilde{\eta}$ onto the column span of $\tilde{X}(P)$ and the residual from this projection, respectively. Note that $\|\eta^\perp\|^2 = J_W(P)$, so the J-statistic is directly informative about the magnitude of this term, while $\hat{\eta} = \tilde{X}(P)(\theta_W(P) - \theta)$ governs the difference between the true and pseudo-true parameter values. Intuitively, $\eta^\perp$ is the detectable component of the misspecification vector η, which has no effect on the bias of the pseudo-true value but governs the J-statistic. Analogously, $\hat{\eta}$ is the undetectable component, which governs the bias but has no effect on the J-statistic. The overall degree of misspecification reflects the sum of these terms, $\|\eta\|_W^2 = \|\eta^\perp\|^2 + \|\hat{\eta}\|^2$. Figure 1 visualizes this decomposition in a case where θ is scalar.

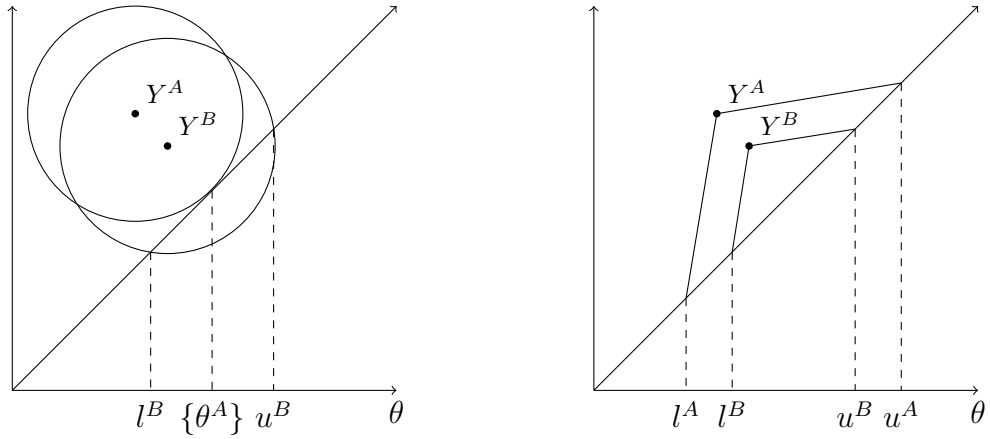

Figure 2: Intervals for θ using the norm-bounding approach and the rotation invariant prior approach. When dataset A or B is observed, the identified set or confidence interval for θ is given by $[l^A, u^A]$ or $[l^B, u^B]$, respectively, where in the left panel $l^A = u^A = \theta^A$.

Equipped with this decomposition, note that for

$$\hat{\eta}(\theta, P) = W(P)^{\frac{1}{2}} X(P)(\theta_W(P) - \theta)$$

the value of $\hat{\eta}$ implied by θ, we can write

$$\Theta_I(P, d) = \left\{\theta : \|\hat{\eta}(\theta, P)\|^2 + J_W(P) \leq d^2\right\}.$$

which implies an identified set for $v'\theta$ equal to

$$\left[v'\theta_W(P) \pm \sqrt{d^2 - J_W(P)}\sqrt{v'H(P)^{-1}v}\right].$$

Hence, the bounds of the identified set correspond to values of θ which spend the misspecification "budget" $\|\eta\|_W^2 \leq d^2 - J_W(P)$ obtained by subtracting the J-statistic from the a-priori upper bound d^2. As the degree of detectable misspecification becomes more severe, in the sense that the J-statistic grows larger, the length of the identified set shrinks. The first panel of Figure 2 illustrates this, again focusing on the case where θ is scalar. Here we hold d, $X(P)$, and $W(P)$ fixed but consider two possible values $Y(P)$, Y^A and Y^B, where Y^A implies a larger J-statistic. The identified set for θ is larger for Y^B than Y^A. Indeed, in this example the J-statistic at Y^A is exactly equal to d, so the identified set collapses to the pseudo-true parameter value.

The comparative statics of our proposed confidence interval (7) are quite different. The proof for the validity of this interval rests on the fact that when $\tilde{\eta}$ is rotation invariant and θ is scalar, the distribution of $\|\hat{\eta}\|/\|\eta^\perp\|$ corresponds exactly to a t-distribution with $k - p$ degrees of freedom. Consistent with this observation, we can

re-write the confidence interval (7) as

$$CI(P) = \left\{\theta : \|\hat{\eta}(\theta, P)\| \leq t^*_{k-p,1-\beta}\|\eta^\perp\|\right\},$$

for $\hat{\eta}(\theta, P) = \tilde{X}(P)(\theta_W(P) - \theta)$. As the second panel of Figure 2 illustrates, this width of this interval is increasing in the size of the J-statistic, with a wider interval for Y^A than for Y^B. This seems like a potentially appealing property for uncertainty summaries in settings where researchers are concerned about misspecification.

6 Conclusion

As highlighted in the literature studying inference on pseudo-true values, true and pseudo-true values generally differ in misspecified models. Consistent with this observation, we show that pseudo-true values do not in general provide a satisfactory data summary for decision-makers whose loss depends on the true parameter value. We also show, however, that for a class of Bayesian priors motivated by minimum distance methods, it is possible to construct confidence intervals that have correct average coverage under the prior without any ex-ante bound on the magnitude of misspecification.

The class of priors we study infers the "shape" of beliefs about model misspecification from the weighting matrix used in minimum distance estimation. This appears to be consistent with the way in which some researchers already choose their weighting matrices.[3] For such researchers, the intervals we suggest offer a natural way to account for the possibility of model misspecification, and are thus a natural complement to standard minimum distance approaches.

[3]For instance, Benhabib et al. (2019) write "The weighting matrix W in the baseline is a diagonal matrix with identical weights for all but the last moment of both the wealth distribution and the mobility moments, which are overweighted (ten times), according to the prior that matching the tail of the distribution is a fundamental objective of our exercise."

References

Alquier, P., J. Ridgway, and N. Chopin (2016): "On the properties of variational approximations of Gibbs posteriors," *Journal of Machine Learning Research*, 17, 1–41.

Andrews, D. W. K. and S. Kwon (2023): "Misspecified Moment Inequality Models: Inference and Diagnostics," *The Review of Economic Studies*, 2674.

Andrews, I. and J. M. Shapiro (2021): "A Model of Scientific Communication," *Econometrica*, 89, 2117–2142.

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

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

Benhabib, J., A. Bisin, and M. Luo (2019): "Wealth Distribution and Social Mobility in the US: A Quantitative Approach," *American Economic Review*, 109, 1623–1647.

Bissiri, P. G., C. C. Holmes, and S. G. Walker (2016): "A general framework for updating belief distributions," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78, 1103–1130.

Catoni, O. (2007): "Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning," *Lecture OPTnotes-Monograph Series*, 56, i–163.

Chernozhukov, V. and H. Hong (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115, 293–346.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly Exogenous," *The Review of Economics and Statistics*, 94, 260–272.

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

Hansen, B. E. and S. Lee (2021): "Inference for Iterated GMM Under Misspecification," *Econometrica*, 89, 1419–1447.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Manski, C. F. (2021): "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827–2853.

Manski, C. F. and J. V. Pepper (2018): "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *The Review of Economics and Statistics*, 100, 232–244.

Martin, R. and N. Syring (2022): "Chapter 1 - Direct Gibbs posterior inference on risk minimizers: Construction, concentration, and calibration," in *Handbook of Statistics*, ed. by A. S. R. Srinivasa Rao, G. A. Young, and C. R. Rao, Elsevier, vol. 47 of *Advancements in Bayesian Methods and Implementation*, 1–41.

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

Muirhead, R. J. (1982): *Aspects of Multivariate Statistical Theory*, Wiley Series in Probability and Statistics, Wiley, 1 ed.

Müller, U. K. (2013): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix," *Econometrica*, 81, 1805–1849.

Rambachan, A. and J. Roth (2023): "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, 90, 2555–2591.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer Series in Statistics, New York, NY: Springer.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

7 Proofs

Proof of Lemma 1 $Q_W(\cdot|P)$ can be expressed as a function of $(X(P), W(P), Y(P))$, so by Assumption 1,

$$\begin{aligned}\pi(\theta \mid P) &= \pi(\theta \mid Q_W(\cdot; P), X(P), W(P)) \\ &= \pi(\theta \mid X(P), W(P), Y(P))\end{aligned}$$

as we aimed to show. □

Proof of Lemma 2 For any parameter value θ, Lemma 1 and Assumption 1 imply that

$$\pi(\theta|P) \propto \pi_\theta(\theta)\pi_\eta(g(\theta;P)|\theta) \propto h(Q_W(\theta;P), W(P), X(P), \theta).$$

Hence, we see that

$$\pi_\eta(g(\theta;P)|\theta) \propto f(Q_W(\theta;P)|W(P), X(P), \theta) := \frac{h(Q_W(\theta;P), W(P), X(P), \theta)}{\pi_\theta(\theta)}$$

where integrability of f follows from the fact that π_η is a probability density. □

Proof of Proposition 1 Immediate from Equation (4) and Lemma 2. □

Proof of Proposition 2 Note that for the claim to hold, it is necessary and sufficient that for $\tilde{W}(P) = X(P)'W(P)X(P)$ and

$$\tilde{B}_\varepsilon(\theta_W(P)) = \left\{\theta : \|\theta_W(P) - \theta\|_{\tilde{W}(P)} < \varepsilon\right\},$$

we have that for all $\varepsilon > 0$,

$$\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\} d\pi_c(\theta \mid P) = \frac{\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}{\int f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta} \to 1.$$

Taking the inverse of this probability (which is possible because the posterior assigns strictly positive mass to neighborhoods of the pseudo-true value) yields

$$1 + \frac{\int 1\{\theta \notin \tilde{B}_\varepsilon(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}, \tag{9}$$

where to prove the result it suffices to show that the second term goes to zero.

To this end, note that for any $a > 1$, continuity of the conditional prior for θ implies that we can re-write the second term of (9) as

$$\frac{\int (1\{\theta \in \tilde{B}_{a\varepsilon}(\theta_W(P)) \setminus \tilde{B}_\varepsilon(\theta_W(P))\} + 1\{\theta \notin \tilde{B}_{a\varepsilon}(\theta_W(P))\}) f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}.$$

Note, however, that we can express the minimum distance objective as

$$Q_W(\theta;P) = J_W(P) + \|\theta - \theta_W(P)\|^2_{\tilde{W}(P)}.$$

Thus, since we have assumed f is non-increasing

$$\frac{\int 1\{\theta \notin \tilde{B}_{a\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta} \leq$$

$$\frac{\int 1\{\theta \notin \tilde{B}_{a\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}(J_W(P)+a^2\varepsilon^2)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}(J_W(P)+\varepsilon^2)\right)\pi_\theta(\theta)d\theta} =$$

$$\frac{f\left(\frac{1}{c}(J_W(P)+a^2\varepsilon^2)\right)}{f\left(\frac{1}{c}(J_W(P)+\varepsilon^2)\right)} \frac{\int 1\{\theta \notin \tilde{B}_{a\varepsilon}(\theta_W(P))\}\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_{\varepsilon}(\theta_W(P))\}\pi_\theta(\theta)d\theta},$$

where the first term converges to zero as $c \to 0$ by Assumption 2, while the second doesn't depend on c. Hence,

$$\lim_{c\to 0} \frac{\int 1\{\theta \notin \tilde{B}_{a\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta} = 0.$$

Note, next, that

$$\frac{\int 1\{\theta \in \tilde{B}_{a\varepsilon}(\theta_W(P)) \setminus \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta} \leq$$

$$\frac{\int 1\{\theta \in \tilde{B}_{a\varepsilon}(\theta_W(P)) \setminus \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}(J_W(P)+\varepsilon^2)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}(J_W(P)+\varepsilon^2)\right)\pi_\theta(\theta)d\theta} =$$

$$\frac{\int 1\{\theta \in \tilde{B}_{a\varepsilon}(\theta_W(P)) \setminus \tilde{B}_{\varepsilon}(\theta_W(P))\}\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_{\varepsilon}(\theta_W(P))\}\pi_\theta(\theta)d\theta},$$

where the last expression goes to zero as we take $a \to 1$. Together with our earlier argument this implies that

$$\lim_{c\to 0} \frac{\int 1\{\theta \notin \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta}{\int 1\{\theta \in \tilde{B}_{\varepsilon}(\theta_W(P))\} f\left(\frac{1}{c}Q_W(\theta;P)\right)\pi_\theta(\theta)d\theta} = 0, \tag{10}$$

and so completes the proof. □

Proof of Proposition 3 Since the loss is uniformly bounded, Proposition 2, together with the dominated convergence theorem, implies that

$$\lim_{c\to 0} \int L(a,\theta) d\pi_c(\theta|P) = L(a,\theta_W(P)) \text{ for all } a \in \mathcal{A}.$$

Our assumptions that L is Lipschitz and $\mathcal{A}$ is compact implies that this convergence is uniform on $\mathcal{A}$, $\lim_{c\to 0} \| \int L(\cdot,\theta) d\pi_c(\theta|P) - L(\cdot,\theta_W(P))\|_\infty = 0$. The result is then immediate from the argmax continuous mapping theorem (Theorem 3.2.2 of van der Vaart and Wellner (1996)) . □

Proof of Proposition 4 For this result we consider priors of the restricted form

$$\pi^{\phi}_{\eta,c}(\eta|\theta) = (1-\phi)\,\pi_{\eta,c}(\eta|\theta) + \phi\pi^*(\eta|\theta).$$

The corresponding posterior is

$$\pi_c^{\phi}(\theta|P) = \frac{\pi_\theta(\theta)\,\pi_{\eta,c}^{\phi}(g(\theta;P)\,|\theta)}{\int \pi_\theta(\theta)\,\pi_{\eta,c}^{\phi}(g(\theta;P)\,|\theta)\,d\theta} =$$

$$\frac{\pi_\theta(\theta)\left((1-\phi)\,\pi_{\eta,c}(g(\theta;P)\,|\theta) + \phi\pi^*(g(\theta;P)\,|\theta)\right)}{\int \pi_\theta(\theta)\left((1-\phi)\,\pi_{\eta,c}(g(\theta;P)\,|\theta) + \phi\pi^*(g(\theta;P)\,|\theta)\right)d\theta}.$$

When $J_W(P) > 0$ we have that $\pi_{\eta,c}(g(\theta;P)\,|\theta) \to 0$ as $c \to 0$ for all θ. Hence, for each θ,

$$\lim_{c\to 0} \pi_\theta(\theta)\left((1-\phi)\,\pi_{\eta,c}(g(\theta;P)\,|\theta) + \phi\pi^*(g(\theta;P)\,|\theta)\right) = \phi\pi_\theta(\theta)\,\pi^*(g(\theta;P)\,|\theta).$$

Moreover, since $\pi_{\eta,c}(g(\theta;P)\,|\theta) \leq \pi_{\eta,c}(g(\theta_W(P);P)\,|\theta)$ by definition, the dominated convergence theorem implies that

$$\int \pi_\theta(\theta)\left((1-\phi)\,\pi_{\eta,c}(g(P,\theta)\,|\theta) + \phi\pi^*(g(P,\theta)\,|\theta)\right)d\theta \to$$

$$\phi\int \pi_\theta(\theta)\,\pi^*(g(P,\theta)\,|\theta)\,d\theta.$$

Therefore the posterior fails to concentrate around the pseudo-true parameter:

$$\lim_{c\to 0}\pi_c^{\phi}(\theta|P) = \frac{\pi_\theta(\theta)\,\pi^*(g(P,\theta)\,|\theta)}{\int \pi_\theta(\theta)\,\pi^*(g(P,\theta)\,|\theta)\,d\theta}.$$

Now we prove the result for $J_W(P) = 0$. Define $w_{1,c} + w_{2,c} = 1$ as

$$w_{1,c} = \frac{(1-\phi)\int \pi_{\eta,c}(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta}{(1-\phi)\int \pi_{\eta,c}(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta + \phi\int \pi^*(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta}$$

and

$$w_{2,c} = \frac{\phi\int \pi^*(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta}{(1-\phi)\int \pi_{\eta,c}(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta + \phi\int \pi^*(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta}.$$

Then we can write the posterior as the following mixture of familiar posterior densities, where $w_{1,c}$ and $w_{2,c}$ are the mixture probabilities,

$$\begin{aligned}
\pi_c^{\phi}(\theta|P) &= \frac{\pi_\theta(\theta)\left((1-\phi)\pi_{\eta,c}(g(\theta;P)|\theta) + \phi\pi^*(g(\theta;P)|\theta)\right))}{\int \pi_\theta(\theta)\left((1-\phi)\pi_{\eta,c}(g(\theta;P)|\theta) + \phi\pi^*(g(\theta;P)|\theta)\right))\,d\theta} \\
&= \frac{(1-\phi)\pi_{\eta,c}(g(\theta;P)|\theta)\pi_\theta(\theta) + \phi\pi^*(g(\theta;P)|\theta))\pi_\theta(\theta)}{(1-\phi)\int \pi_{\eta,c}(g(\theta;P)|\theta)\pi_\theta(\theta)d\theta + \phi\int \pi^*(g(\theta;P)|\theta))\pi_\theta(\theta)d\theta} \\
&= \frac{(1-\phi)\left[\int \pi_{\eta,c}(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta\right]\pi_c(\theta \mid P) + \phi\left[\int \pi^*(g(\theta;P) \mid \theta)\,\pi_\theta(\theta)d\theta\right]\pi^*(\theta \mid P)}{(1-\phi)\int \pi_{\eta,c}(g(\theta;P)|\theta)\pi_\theta(\theta)d\theta + \phi\int \pi^*(g(\theta;P)|\theta))\pi_\theta(\theta)d\theta} \\
&= w_{1,c}\pi_c(\theta \mid P) + w_{2,c}\pi^*(\theta \mid P).
\end{aligned}$$

Proposition 2 states that the first mixture component concentrates around the pseudo-true value as $c \to 0$. Meanwhile, the second mixture component is constant with respect to c. Therefore, we are interested in the limiting behavior of the mixture probabilities as $c \to 0$ when $J_W(P) = 0$. We next show how the numerator of $w_{1,c}$ converges to $\pi_\theta(\theta_W(P))$ as $c \to 0$. Note that when $J_W(P) = 0$ we can write the minimum distance objective as $Q_W(\theta; P) = \|\theta - \theta_W(P)\|_{\tilde{W}(P)}$ so the integral found in $w_{1,c}$ and $w_{2,c}$ can be written as

$$\int \pi_{\eta,c}(g(\theta; P) \mid \theta)\pi_\theta(\theta)d\theta = \int \pi_\theta(\theta) \frac{f\left(\frac{1}{c}\|\theta - \theta_W(P)\|^2_{\tilde{W}(P)}\right)}{\int f\left(\frac{1}{c}\|\tilde{\theta} - \theta_W(P)\|^2_{\tilde{W}(P)}\right) d\tilde{\theta}} d\theta.$$

Make the substitution $\theta \mapsto \theta_W(P) - \sqrt{c}\theta$ in the outer integral and $\tilde{\theta} \mapsto \theta_W(P) - \sqrt{c}\tilde{\theta}$ in the normalizing constant as follows,

$$\int \pi_\theta(\theta) \frac{f\left(\frac{1}{c}\|\theta - \theta_W(P)\|^2_{\tilde{W}(P)}\right)}{\int f\left(\frac{1}{c}\|\tilde{\theta} - \theta_W(P)\|^2_{\tilde{W}(P)}\right) d\tilde{\theta}} d\theta =$$

$$\int c^{-1/2}\pi_\theta(\theta_W(P) - \sqrt{c}\theta) \frac{f\left(\|\theta\|^2_{\tilde{W}(P)}\right)}{\int c^{-1/2} f\left(\|\tilde{\theta}\|^2_{\tilde{W}(P)}\right) d\tilde{\theta}} d\theta =$$

$$\int \pi_\theta(\theta_W(P) - \sqrt{c}\theta) \frac{f\left(\|\theta\|^2_{\tilde{W}(P)}\right)}{\int f\left(\|\tilde{\theta}\|^2_{\tilde{W}(P)}\right) d\tilde{\theta}} d\theta$$

Because π_θ is assumed to be continuous and full-support, there exists some finite M such that $\pi_\theta(\theta) \leq M$ for all θ in a neighborhood of $\theta_W(P)$. It follows that

$$\lim_{c \to 0} \int \pi_\theta(\theta_W(P) - \sqrt{c}\theta) \frac{f\left(\|\theta\|^2_{\tilde{W}(P)}\right)}{\int f\left(\|\theta\|^2_{\tilde{W}(P)}\right) d\theta} d\theta = \pi_\theta(\theta_W(P)).$$

Therefore, when we take $c \to 0$, the mixture probabilities converge to the following limits:

$$w_{1,c} \to w_1^* = \frac{(1-\phi)\pi_\theta(\theta_W(P))}{(1-\phi)\pi_\theta(\theta_W(P)) + \phi \int \pi^* \left(g(\theta; P) \mid \theta\right) \pi_\theta(\theta) d\theta}$$

and

$$w_{2,c} \to (1 - w_1^*) = \frac{\phi \int \pi^* \left(g(\theta; P) \mid \theta\right) \pi_\theta(\theta) d\theta}{(1-\phi)\pi_\theta(\theta_W(P)) + \phi \int \pi^* \left(g(\theta; P) \mid \theta\right) \pi_\theta(\theta) d\theta}.$$

Thus as we take $c \to 0$,

$$\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\}\pi_c^\phi(\theta|P)d\theta =$$

$$w_{1,c}\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\}\pi_c(\theta|P)d\theta + w_{2,c}\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\}\pi^*(\theta|P)d\theta \rightarrow$$

$$\rightarrow w_1^* + (1-w_1^*)\int 1\{\theta \in \tilde{B}_\varepsilon(\theta_W(P))\}\pi^*(\theta|P)d\theta.$$

Since the mixture posterior $\pi^*(\theta|P)$ is assumed to be continuous, taking $\varepsilon \rightarrow 0$ sends the second term to zero. Therefore, the posterior concentrates around a point mass at the pseudo-true value with probability w_1^*, where $w_1^* \rightarrow 1$ as $\phi \rightarrow 0$. □

Proof of Proposition 5 Classic results in statistics, e.g. Muirhead (1982), Chapter 1.5, imply the result for the case of a single regressor. For completeness, we prove the result for the general case.

The confidence interval (7) corresponds to the set of values for $v'\theta$ where the test statistic

$$\frac{v'\theta_W(P) - v'\theta}{\sqrt{\frac{J_W(P)}{k-p}v'H(P)^{-1}v}} \tag{11}$$

has absolute value less than $t^*_{k-p,1-\frac{\beta}{2}}$. Hence, if we can show that (11) follows a t_{k-p} distribution under all priors consistent with Assumption 1, the result is immediate.

Note that (11) is equal to the t-statistic from regression (8) in the text,

$$\frac{v'\theta_W(P) - v'\theta}{\sqrt{\frac{J_W(P)}{k-p}v'H(P)^{-1}v}} = \frac{v'(\tilde{X}(P)'\tilde{X}(P))^{-1}\tilde{X}(P)'\tilde{\eta}}{\sqrt{\frac{\tilde{\eta}'(I-M(P))\tilde{\eta}}{k-p}v'H(P)^{-1}v}} \tag{12}$$

Specifically, $\frac{\tilde{\eta}'(I-M(P))\tilde{\eta}}{k-p}$ corresponds to the unbiased variance estimate, so the denominator $\sqrt{\frac{\tilde{\eta}'(I-M(P))\tilde{\eta}}{k-p}v'H(P)^{-1}v}$ corresponds to the homoskedastic standard error. Note, moreover, that the t-statistic is scale-invariant in the error, so (12) is equal to

$$\frac{v'(\tilde{X}(P)'\tilde{X}(P))^{-1}\tilde{X}(P)'\frac{\tilde{\eta}}{\|\eta\|}}{\sqrt{\frac{\frac{\tilde{\eta}}{\|\eta\|}'(I-M(P))\frac{\tilde{\eta}}{\|\eta\|}}{k-p}v'H(P)^{-1}v}}.$$

Proposition 1 implies that $\frac{\tilde{\eta}}{\|\eta\|}$ is uniformly distributed on the unit sphere under any prior π satisfying Assumption 1. However, this exactly the distribution of $\frac{Z}{\|Z\|}$ for $Z \sim N(0,I)$. It follows that under π, (12) has the same distribution as

$$\frac{v'(\tilde{X}(P)'\tilde{X}(P))^{-1}\tilde{X}(P)'\frac{Z}{\|Z\|}}{\sqrt{\frac{\frac{Z}{\|Z\|}'(I-M(P))\frac{Z}{\|Z\|}}{k-p}v'H(P)^{-1}v}} = \frac{v'(\tilde{X}(P)'\tilde{X}(P))^{-1}\tilde{X}(P)'Z}{\sqrt{\frac{Z'(I-M(P))Z}{k-p}v'H(P)^{-1}v}},$$

where we have again used scale invariance of the t-statistic. However, the last expression is the t-statistic from the regression

$$\tilde{Y}(P) = \tilde{X}(P)\theta + Z,$$

which is well-known to be t_{k-p} distributed, completing the proof. $\square$