Minimizing Sensitivity to Model Misspecification

Stéphane Bonhomme
University of Chicago

Martin Weidner
University College London

Chicago
April 2019

Model misspecification

• Economic models are intended as approximations. Yet econometric inference typically hinges on correct specification.

• For example, economists often use fully-specified models to predict the effect of a policy or another counterfactual.

-It is common practice to "plug in" the estimated parameters in a formula to compute the effect of interest, solely based on the model.

• Our goal is to propose a framework for point estimation and inference when the model is misspecified.

Two illustrations

1. Counterfactual predictions from a structural model

-We revisit the debate between Todd and Wolpin (2006) and Attanasio, Meghir and Santiago (2011), who combine the PROGRESA randomized experiment with structural models of education choice. We estimate a simple model in the spirit of TW, and adjust its predictions against a form of misspecification emphasized by AMS.

2. Average effects in panel data

-Panel data settings: $Y_{i1}, ..., Y_{iT}$, $X_{i1}, ..., X_{iT}$, and latent individual effects A_i. Large literature on misspecification of $\pi(A_i \mid X_i, Y_{i1})$. We propose a robustness approach for fixed T, based on minimizing the impact of deviations in a neighborhood of a parametric reference model (e.g., a correlated random-effects specification).

Setup

• "Large model" correctly specified: Let $f_\theta(y)$ be a model indexed by a (finite or infinite) parameter $\theta \in \Theta$, we observe $Y_1, \ldots, Y_n \sim iid\, f_\theta(y)$.

• Our target parameter is δ_θ, a scalar function of θ. The true value θ_0 is unknown.

• "Small model" may be misspecified: The researcher has a parametric reference model $f_{\theta(\eta)}$, and an estimate $\widehat{\eta}$ of the finite-dimensional parameter $\eta \in \mathcal{B}$.

• We consider θ_0 values in a neighborhood of the reference model,

$$\Gamma_\epsilon = \{(\theta_0, \eta) \in \Theta \times \mathcal{B} : d(\theta_0, \theta(\eta)) \leq \epsilon\},$$

where d is a (squared) distance measure, and ϵ is neighborhood size.

Example: panel data models

- Outcomes $Y_i = (Y_{i1}, \ldots, Y_{iT})$, covariates X_i, individual effects A_i.
- For $\theta = (\beta, \pi)$ the distribution function of Y given X is

$$f_\theta(y \,|\, x) = \int_{\mathcal{A}} g_\beta(y \,|\, a, x) \pi(a \,|\, x) da.$$

- We want to estimate average effects of the form

$$\delta_{\theta_0} = \mathbb{E}_{\theta_0}\left[\Delta(A, X, \beta_0)\right] = \iint_{\mathcal{A}\times\mathcal{X}} \Delta(a, x, \beta_0) \pi_0(a \,|\, x) f_X(x) da dx,$$

- Reference model: correlated random-effects specification $\pi_\gamma(a \,|\, x)$.
- In this setup, $\theta = (\beta, \pi)$, $\eta = (\beta, \gamma)$, and $\theta(\eta) = (\beta, \pi_\gamma)$. We consider

$$d(\theta_0, \theta) = \|\beta_0 - \beta\|_\Omega^2 + 2 \int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right) \pi_0(a) da.$$

Summary of Approach

- Naive estimator $\delta_{\theta(\widehat{\eta})}$ may be biased $\Rightarrow$ how to robustify?

- We caluclate minimax estimator, which minimizes worst-case MSE

$$\mathbb{E}_{\theta_0}\left[\left(\widehat{\delta} - \delta_{\theta_0}\right)^2\right]$$

in the ϵ-neighborhood of the reference model:

$$\Gamma_\epsilon = \{(\theta_0, \eta) \in \Theta \times \mathcal{B} : d(\theta_0, \theta(\eta)) \leq \epsilon\}$$

- Focus on small ϵ (asymptotic: $\epsilon \to 0$ as $n \to \infty$), which allows to approximate the MSE problem, and find a "one-step" adjustment $\Rightarrow$ tractable way to perform sensitivity analysis.

- We also construct confidence intervals which contain δ_{θ_0} asymptotically with pre-specified probability, uniformly on Γ_ϵ.

Summary of Approach (cont.)

- We focus on two cases: parametric misspecification based on a weighted Euclidean norm, and semi-parametric misspecification of functional forms based on Kullback-Leibler.

- We use a calibration approach for ϵ, targeting the probability of a model detection error, as in Hansen and Sargent (2008).

- It is not necessary to estimate the "larger" model. This feature is reminiscent of LM testing.

- The "large" model need not be point-identified.

Literature

• Robustness in statistics and economics: Huber (1981), Hampel *et al.* (1986), Rieder (1994). Focus on data contamination and fully nonparametric deviations. Also Bayesian literature, e.g. Gustafson (2000). Hansen and Sargent's (2001, 2008) robust decision-making.

• Sensitivity analysis, local misspecification and local robustness: Rosenbaum and Rubin (1983), Leamer (1985), Conley *et al.* (2012), Kitamura *et al.* (2013), Andrews *et al.* (2017, 2018), Armstrong and Kolesar (2018). Neyman's orthogonalization: Neyman (1959), Chernozhukov *et al.* (2016). Here: MSE, confidence intervals.

• Partial identification for sensitivity analysis: Chen, Tamer and Torgovitsky (2011), Norets and Tang (2014), Christensen and Connault (2018). Here: local neighborhood of a reference model.

Worst Case MSE Minimization

Remember the setup

- $f_\theta(y)$ is correctly specified
- Target parameter: δ_θ
- Parametric reference model: $f_{\theta(\eta)}$, with estimator $\widehat{\eta}$
- ϵ-neighborhood: $\Gamma_\epsilon = \{(\theta_0, \eta) \in \Theta \times \mathcal{B} : d(\theta_0, \theta(\eta)) \leq \epsilon\}$

Estimates for δ_θ (assume fixed known η first)

- We focus on asymptotically linear estimators:

$$\widehat{\delta} = \delta_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta) + o_P(\epsilon^{\frac{1}{2}}) + o_P(n^{-\frac{1}{2}}),$$

- Assume that $\widehat{\delta}$ is asymptotically unbiased under reference model:

$$\mathbb{E}_{\theta(\eta)} h(Y, \eta) = 0,$$

- Want to construct an estimator of δ_{θ_0} of the form

$$\widehat{\delta}_{h,\eta} = \widehat{\delta}_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta),$$

for a random sample $Y_1, ..., Y_n$ drawn from f_{θ_0}, such that we minimize

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_{h,\eta} - \delta_{\theta_0}\right)^2\right]$$

Worst Case Bias (small ϵ)

- As $\epsilon \to 0$, the bias $\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \left| \mathbb{E}_{\theta_0} \left[\widehat{\delta} - \delta_{\theta_0} \right] \right|$ can be approximated by

$$b_\epsilon(h, \eta) = \epsilon^{1/2} \left\| \nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} \, h(Y, \eta) \, \nabla_\theta \log f_\theta(Y) \right\|_\eta .$$

where for a for any linear map $u : \Theta \to \mathbb{R}$ we define

$$\|u\|_\eta = \lim_{\epsilon \to 0} \|u\|_{\eta,\epsilon} \qquad \|u\|_{\eta,\epsilon} = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \epsilon^{-\frac{1}{2}} \, u'(\theta_0 - \theta(\eta)).$$

Worst Case MSE (small ϵ)

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_{h,\eta} - \delta_{\theta_0} \right)^2 \right] = [b_\epsilon(h, \eta)]^2 + \mathrm{Var}_{\theta(\eta)} \left(\widehat{\delta}_{h,\eta} \right) + o\left(\epsilon + \frac{1}{n} \right)$$

Minimum MMSE Estimator (known η, $\epsilon \to 0$ as $n \to \infty$)

- Define the minimum-MSE function $h_\epsilon^{\mathsf{MMSE}}(y, \eta)$ as

$$\underset{h(\cdot,\eta)}{\operatorname{argmin}} \; \epsilon \left\| \nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} \, h(Y,\eta) \, \nabla_\theta \log f_\theta(Y) \right\|_\eta^2 + \frac{\mathsf{Var}_{\theta(\eta)}(h(Y,\eta))}{n},$$

subject to asymptotic unbiasedness under $f_{\theta(\eta)}$:

$$\mathbb{E}_{\theta(\eta)}(h(Y,\eta)) = 0.$$

- The minimum-MSE estimator is then

$$\widehat{\delta}_\epsilon^{\mathsf{MMSE}} = \delta_{\theta(\eta)} + \frac{1}{n} \sum_{i=1}^{n} h_\epsilon^{\mathsf{MMSE}}(Y_i, \eta).$$

Minimum MMSE Estimator (estimated η, $\epsilon \to 0$ as $n \to \infty$)

- Define the minimum-MSE function $h_\epsilon^{\mathrm{MMSE}}(y, \eta)$ as

$$\underset{h(\cdot,\eta)}{\operatorname{argmin}}\ \epsilon \left\| \nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)}\, h(Y,\eta)\, \nabla_\theta \log f_\theta(Y) \right\|_\eta^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n},$$

subject to: (1) asymptotic unbiasedness under $f_{\theta(\eta)}$,

$$\mathbb{E}_{\theta(\eta)}(h(Y,\eta)) = 0,$$

and (2) local robustness with respect to η,

$$\mathbb{E}_{\theta(\eta)}\, h(Y,\eta)\, \nabla_\eta \log f_{\theta(\eta)}(Y) = \nabla_\eta \delta_{\theta(\eta)}.$$

- Given a preliminary estimator $\widehat{\eta}$, the minimum-MSE estimator is then

$$\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n} \sum_{i=1}^{n} h_\epsilon^{\mathrm{MMSE}}(Y_i, \widehat{\eta}).$$

Locally Robustness Constraint

- Remember:

$$\widehat{\delta} = \delta_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta) + o_P(\epsilon^{\frac{1}{2}}) + o_P(n^{-\frac{1}{2}}),$$

- Locally robustness with respect to η demands that

$$\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{\theta(\eta)} \nabla_\eta h(Y, \eta) = 0,$$

(see e.g. Chernozhukov, Escanciano, Ichimura and Newey 2016)

Using the unbiasedness constraint this can equivalently be written as

$$\mathbb{E}_{\theta(\eta)}\, h(Y, \eta)\, \nabla_\eta \log f_{\theta(\eta)}(Y) = \nabla_\eta \delta_{\theta(\eta)}.$$

Explicit solution in parametric model

- Consider a weighted Euclidean distance

$$d(\theta_0, \theta(\eta)) = \|\theta_0 - \theta(\eta)\|_\Omega^2 = (\theta_0 - \theta(\eta))' \, \Omega \, (\theta_0 - \theta(\eta))$$

- One then finds

$$h_\epsilon^{\mathsf{MMSE}}(y, \eta) = \left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \left[\widetilde{H}_{\theta(\eta)} + (\epsilon n)^{-1} \Omega\right]^{-1} \widetilde{\nabla}_\theta \delta_{\theta(\eta)},$$

where $\widetilde{\nabla}_\theta = \nabla_\theta - H_{\theta(\eta)} G_\eta' H_\eta^{-1} \nabla_\eta$, $G_\eta = \nabla_\eta \theta(\eta)'$,

$$H_\eta = \mathbb{E}_{\theta(\eta)}\left[\nabla_\eta \log f_{\theta(\eta)}(Y)\right]\left[\nabla_\eta \log f_{\theta(\eta)}(Y)\right]',$$
$$\widetilde{H}_{\theta(\eta)} = \mathsf{Var}\left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right] = H_{\theta(\eta)} - H_{\theta(\eta)} G_\eta' H_\eta^{-1} G_\eta H_{\theta(\eta)}.$$

$\Rightarrow$ Tikhonov regularization

Interpretation

• The minimum-MSE estimator interpolates, nonlinearly, between the one-step approximations to the MLE of the "small" model $\theta(\eta)$ (when ϵ is small) and the "large" model θ (when ϵ is large).

• However, we do not assume that the "large" model f_θ is point-identified. The MMSE estimator only updates the estimator in directions where f_θ has identifying power.

Example: OLS vs IV

- Consider the linear regression model

$$\begin{aligned} Y &= X'\beta + U \\ X &= \Pi Z + V, \end{aligned}$$

where $U = \rho' V + \xi$.

- We assume joint Gaussianity, and take $\theta = (\beta, \rho)$, $\eta = \beta$, and $\theta(\eta) = (\beta, 0)$. The target parameter is $\delta_\theta = c'\beta$, and Ω is block-diagonal with blocks Ω_β and Ω_ρ. Π, Σ_V, σ_ξ^2 assumed known.

- We find

$$\begin{aligned} h_\epsilon^{\mathsf{MMSE}}(y, x, z, \beta) = (y - x'\beta) x' \Sigma_X^{-1} c \\ - (y - x'\beta) \left[(x - \Pi z) - \Sigma_V \Sigma_X^{-1} x \right]' \times \\ \left[\Sigma_V - \Sigma_V \Sigma_X^{-1} \Sigma_V + (\epsilon n)^{-1} \Omega_\rho \right]^{-1} \Sigma_V \Sigma_X^{-1} c. \end{aligned}$$

OLS vs IV (cont.)

• When $\epsilon = 0$, the minimum-MSE estimator of $c'\beta$ is the "one-step efficient" adjustment in the direction of the OLS estimator based on

$$h_0^{\text{MMSE}}(y, x, z, \beta) = (y - x'\beta)x'\Sigma_X^{-1}\, c.$$

• As $\epsilon \to \infty$, provided $\Pi\Sigma_Z\Pi'$ is invertible, the adjustment is performed in the direction of the IV estimator, based on

$$\lim_{\epsilon\to\infty} h_\epsilon^{\text{MMSE}}(y, x, z, \beta) = (y - x'\beta)\,[\Pi z]'\left[\Pi\Sigma_Z\Pi'\right]^{-1} c.$$

• For given $\epsilon > 0$ and n, our $\widehat{\delta}_\epsilon^{\text{MMSE}}$ remains well-defined when $\Pi\Sigma_Z\Pi'$ is singular.

Confidence intervals

• Let $\widehat{\delta}$ be an asymptotically linear estimator with influence function $h(\cdot, \eta)$. Let $\widehat{\sigma}_h^2$ be the sample variance of $h(Y_1, \widehat{\eta}), \ldots, h(Y_n, \widehat{\eta})$.

• Define the following interval, for any confidence level $\mu \in (0, 1)$,

$$CI_\epsilon(1-\mu) = \left[\widehat{\delta} \pm \left(b_\epsilon\left(h, \widehat{\eta}\right) + \frac{\widehat{\sigma}_h}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\mu}{2}\right)\right)\right].$$

• $CI_\epsilon(1-\mu)$ contains δ_{θ_0} with probability approaching $1-\mu$ as n tends to infinity and ϵn tends to a constant, uniformly on Γ_ϵ; that is, both under correct specification and under local misspecification of the reference model.

• Such CI's are "honest" since they take into account the bias of the estimator (e.g., Armstrong and Kolesar, 2016).

Choice of ϵ

- We follow a calibration approach for ϵ as Hansen and Sargent (2008), and target the probability of a model detection error.

- For a fixed probability $p \in (0, 1)$, we choose $\epsilon = \epsilon(p)$ such that

$$\inf_{\theta_0 \in \Gamma_\epsilon(\eta)} \Pr_{\theta(\eta)} \left(\sum_{i=1}^{n} \log \left(\frac{f_{\theta_0}(Y_i)}{f_{\theta(\widehat{\eta})}(Y_i)} \right) > 0 \right) = p + o(1).$$

- For some θ_0 in $\Gamma_\epsilon(\eta)$, the probability of incorrectly detecting that θ_0 is more likely to have generated the data than $\theta(\eta)$ is approximately equal to p.

- Achieving a lower p requires setting a higher ϵ.

Choice of ϵ (cont.)

- For fixed p, as n tend to infinity ϵn tends to a constant and we have

$$\epsilon = \frac{4\left(\Phi^{-1}(p)\right)^2}{n \cdot \lambda_{max}(\Omega^{-\frac{1}{2}}\widetilde{H}_{\theta(\eta)}\Omega^{-\frac{1}{2}})} + o\left(n^{-1}\right).$$

- *"[such] neighborhoods make eminent sense, since the standard goodness-of-fit tests are just able to detect deviations of this order. Larger deviations should be taken care of by diagnostic and modeling, while smaller ones are difficult to detect and should be covered by robustness"* (Huber and Ronchetti, 2009).

Panel Data (semi-parametric model)

Remember

- $\theta = (\beta, \pi)$ and $f_\theta(y \mid x) = \int_{\mathcal{A}} g_\beta(y \mid a, x)\pi(a \mid x)da$.
- We want to estimate $\delta_{\theta_0} = \mathbb{E}_{\theta_0}[\Delta(A, X, \beta_0)]$
- Reference model: $\pi_\gamma(a \mid x)$, $\eta = (\beta, \gamma)$
- Distance measure: $d(\theta_0, \theta) = \|\beta_0 - \beta\|^2_\Omega + 2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right)\pi_0(a)da$.

Worst Case Bias:

- As $\epsilon \to 0$ the bias $\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \left|\mathbb{E}_{\theta_0}\left[\widehat{\delta} - \delta_{\theta_0}\right]\right|$ can be approximated by

$$b_\epsilon(h, \beta, \gamma) = \epsilon^{1/2}\sqrt{\mathrm{Var}_\gamma\left(\Delta(A, \beta) - \mathbb{E}_\beta\left[h(Y) \mid A\right]\right)}.$$

Minimum-MSE in the panel data case

- For fixed ϵ and n, the linear system

$$\mathbb{E}_{\beta,\gamma}\left[\mathbb{E}_{\beta}(h_{\epsilon}^{\mathsf{MMSE}}(Y,\beta,\gamma)\,|\,A)\mid y\right]+(\epsilon n)^{-1}h_{\epsilon}^{\mathsf{MMSE}}(y,\beta,\gamma)$$
$$=\mathbb{E}_{\beta,\gamma}\left[\Delta(A,\beta)\mid y\right]-\mathbb{E}_{\gamma}\Delta(A,\beta)$$

has a unique solution irrespective of whether the following operator is injective or its inverse is ill-posed,

$$\mathbb{H}\,:\,h(y)\mapsto\mathbb{E}_{\beta,\gamma}\left[\Delta(A,\beta)-\mathbb{E}_{\beta}(h(Y)\,|\,A)\mid y\right].$$

- This covers cases where the parameter π is not point-identified, or π is irregularly point-identified.

- In the latter case, $(\epsilon n)^{-1}$ acts as a calibrated (Tikhonov) regularization parameter.

Panel data estimators: RE and EB

• Consider the *random-effects* estimator

$$\widehat{\delta}^{\mathsf{RE}} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{A}} \Delta\left(a, X_i, \widehat{\beta}\right) \pi_{\widehat{\gamma}}(a \,|\, X_i)\, da,$$

where $(\widehat{\beta}, \widehat{\gamma})$ is the maximum likelihood estimator of (β, γ) based on the reference model. Consider also the *empirical Bayes* estimator

$$\widehat{\delta}^{\mathsf{EB}} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{A}} \Delta(a, X_i, \widehat{\beta})\, p_{\widehat{\beta},\widehat{\gamma}}(a \,|\, Y_i, X_i)\, da,$$

where $p_{\widehat{\beta},\widehat{\gamma}}(a \,|\, Y_i, X_i)$ denotes the posterior distribution of individual effects A_i given (Y_i, X_i) implied by $g_{\widehat{\beta}}$ and $\pi_{\widehat{\gamma}}$.

• Both RE and EB are consistent for fixed T as n tends to infinity under correct specification. However, misspecification of π_γ makes RE and EB fixed-T inconsistent.

Revisiting the dynamic probit model

• We perform numerical simulations based on the following dynamic panel data probit model with individual effects,

$$Y_{it} = \mathbf{1}\left\{\beta_0 Y_{i,t-1} + A_i + U_{it} \geq 0\right\}, \quad t = 1, ..., T,$$

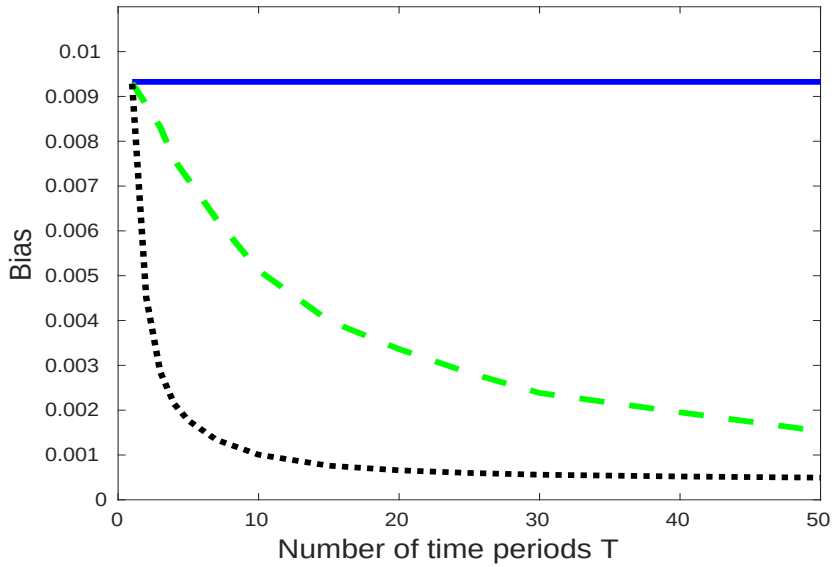
where $U_{i1}, ..., U_{iT}$ are i.i.d. standard normal, independent of Y_{i0} and A_i. We take the density $\pi = \pi_{\mu,\sigma}$ of A_i given Y_{i0} to be Gaussian with mean $\mu_0 + \mu_1 Y_{i0}$ and variance σ^2. Here we treat $\mu_0 = -.25, \mu_1 = .5, \sigma = .8$ as known. $\beta_0 = .5$. Y_{i0} is Bernoulli(.5).

• We focus on the average *state dependence* effect

$$\delta_{\theta_0} = \mathbb{E}_{\pi_0}\left[\Phi(\beta_0 + A_i) - \Phi(A_i)\right],$$

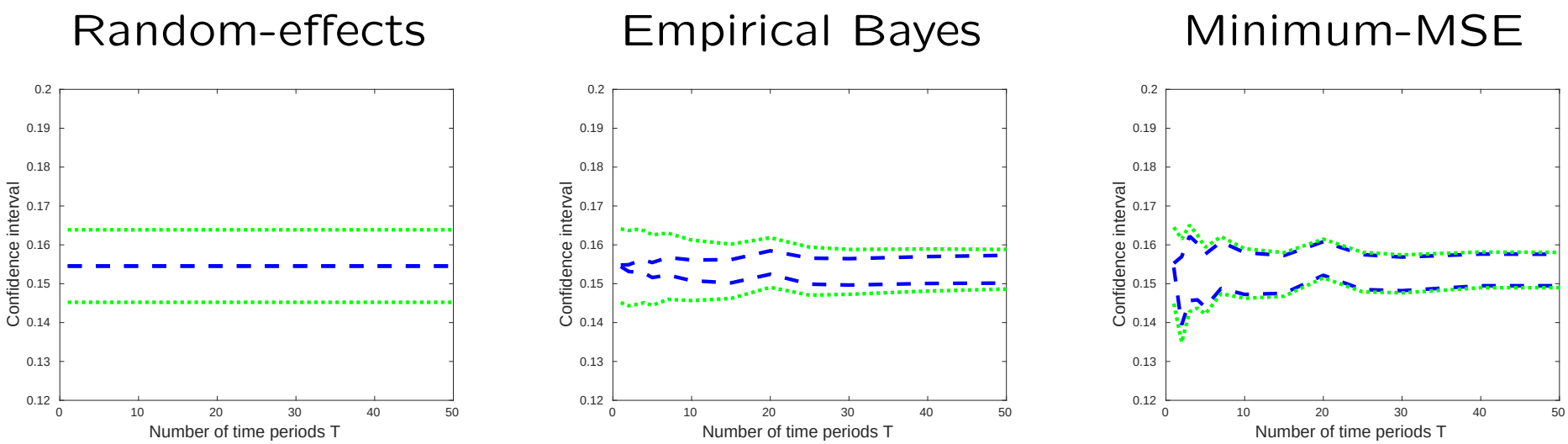
and we consider the RE, EB, and minimum-MSE estimators. The latter is computed based on simple matrix calculations (see paper).

Bias of different estimators of the average state dependence effect in the dynamic probit model ($p = .01$)



Notes: Bias b_ϵ for different panel length T. The solid line corresponds to the random-effects estimator $\widehat{\delta}^{\text{RE}}$, the dashed line to the empirical Bayes estimator $\widehat{\delta}^{\text{EB}}$, and the dotted line to the minimum-MSE estimator $\widehat{\delta}_\epsilon^{\text{MMSE}}$. ϵ is calibrated to a detection error probability $p = .01$ when $n = 500$.

Confidence intervals of the average state dependence effect in the dynamic probit model ($p = .01$)



Notes: 95%-confidence intervals for the average state dependence effect, based on three estimators. Dashed lines correspond to confidence intervals based on correct specification, dotted lines to the ones allowing for local misspecification. $n = 500$. ϵ *is calibrated to a detection error probability* $p = .01$.

Monte Carlo simulation of the average state dependence effect δ and autoregressive parameter β in the dynamic probit model, log-normal specification for A_i

T	5	10	20	50	5	10	20	50
	Bias				MSE ($\times$1000)			
	Average state dependence δ							
Random-effects	-.067	-.065	-.047	-.033	4.73	4.31	2.27	1.11
Empirical Bayes	-.065	-.059	-.035	-.016	4.43	3.57	1.30	.278
Linear probability	-.299	-.124	-.052	-.011	90.0	15.7	2.79	.171
Minimum-MSE ($p = .01$)	-.021	-.005	.000	.001	1.14	.408	.163	.075
Minimum-MSE ($p = .10^{-10}$)	-.005	.002	.003	.002	1.02	.454	.187	.086
	Autoregressive parameter β							
Maximum likelihood	-.154	-.146	-.085	-.038	26.3	22.8	7.82	1.72
Minimum-MSE ($p = .01$)	-.038	.006	.012	.008	8.02	3.51	1.57	.691
Minimum-MSE ($p = .10^{-10}$)	.005	.025	.019	.011	8.78	4.62	1.89	.781

Notes: $n = 500$, *results for* 1000 *simulations.* $2KL(\pi_0, \pi) = 1.52$. ϵ *is* .04 *for* $p = .01$, .32 *for* $p = 10^{-10}$.

Conclusions ("Minimizing Sensitivity to Model Misspecification")

- We propose an approach to estimation and inference in the presence of model misspecification.

- Our minimax mean squared error rule consists of a one-step adjustment, which is motivated by both robustness and efficiency.

- Implementing our estimators and confidence intervals does not require estimating a larger model.

- Our adjustments remain valid when the identification of the "large" model is irregular or point-identification fails.

Follow-up paper: "Posterior Average Effects" (work in progress)

Summary ("Posterior Average Effects")

- A posterior estimator of δ_θ is (using panel example here)

$$\widehat{\delta}^{\mathsf{P}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_{\theta(\widehat{\eta})}}\left[\Delta(A, X, \widehat{\beta}) \,\middle|\, Y = Y_i, X = X_i\right].$$

- How can this estimator be justified from the perspective of misspecification robustness?

- We show that $\widehat{\delta}^{\mathsf{P}}$ minimizes worst case "bias":

$$\left|\mathbb{E}_{\theta_0}\left(\widehat{\delta} - \delta_{\theta_0}\right)\right|$$

over an appropriately defined ϵ-neighborhood $\widetilde{\Gamma}_\epsilon$ of the reference model.

- This optimality result of $\widehat{\delta}^{\mathsf{P}}$ again requires small ϵ (limit $\epsilon \to 0$), but we also find that $\widehat{\delta}^{\mathsf{P}}$ is optimal up to a factor 2 for finite ϵ.

Model and parameter of interest

- Consider

$$Y = \phi_\beta(U, X),$$

where (Y, X) is observed to the researcher and U is unobserved.

- We wish to estimate an average effect: $\overline{\delta} = \mathbb{E}\left[\delta_\beta(U, X)\right]$

- Model based estimator: (based on parametric model $f_\sigma(U|X)$)

$$\widehat{\delta}^{\mathsf{M}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_{\widehat{\sigma}}}\left[\delta_{\widehat{\beta}}(U, X) \,\middle|\, X = X_i\right].$$

- Posterior estimator:

$$\widehat{\delta}^{\mathsf{P}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{p_{\widehat{\beta}}(f_{\widehat{\sigma}})}\left[\delta_{\widehat{\beta}}(U, X) \,\middle|\, Y = Y_i, X = X_i\right].$$

Estimators and bias

- We focus on estimators of $\overline{\delta}$ that satisfy, for a scalar function h,

$$\widehat{\delta}_h = \frac{1}{n}\sum_{i=1}^{n} h_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i) + o_p(1),$$

and are asymptotically unbiased under the reference distribution, i.e.

$$\mathbb{E}_{f_{\sigma_*}}[h_{\beta,\sigma_*}(Y, X) - \delta_\beta(U, X)] = 0.$$

- Given an estimator $\widehat{\delta}_h$, we define its ϵ-worst-case bias as

$$b_\epsilon(h) = \sup_{f_0 \in \Gamma_\epsilon} \left| \mathbb{E}_{f_0}[h_{\beta,\sigma_*}(Y, X)] - \mathbb{E}_{f_0}[\delta_\beta(U, X)] \right|.$$

where

$$\Gamma_\epsilon = \left\{ f_0 : d(f_0, f_{\sigma_*}) \leq \epsilon,\ \mathbb{E}_{f_0}[\Psi_{\beta,\sigma_*}(Y, X)] = 0 \right\}.$$

Local bias

• Our first result is a local characterization of the bias $b_\epsilon(h)$ of estimators $\widehat{\delta}_h$. We show that, as $\epsilon \to 0$ and under regularity conditions,

$$b_\epsilon(h) = \epsilon^{\frac{1}{2}} \left\{ \mathbb{E}_{f_{\sigma_*}} \left[\left(h_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) - \lambda' \Psi_{\beta,\sigma_*}(Y,X) \right)^2 \right] \right\}^{\frac{1}{2}} + o(\epsilon^{\frac{1}{2}}),$$

where

$$\lambda = \left\{ \mathbb{E}_{f_{\sigma_*}} \left[\Psi_{\beta,\sigma_*} \Psi'_{\beta,\sigma_*} \right] \right\}^{-1} \mathbb{E}_{f_{\sigma_*}} \left[\left(h_{\beta,\sigma_*} - \delta_\beta \right) \Psi_{\beta,\sigma_*} \right].$$

• Hence, to first order, $b_\epsilon(h)$ is minimum whenever

$$\mathbb{E}_{f_{\sigma_*}} \left[\left(h_{\beta,\sigma_*}(Y,X) - \mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X) \,|\, Y,X] - \lambda' \Psi_{\beta,\sigma_*}(Y,X) \right)^2 \right]$$

is minimum. It is easy to see that $h^{\mathsf{P}}_{\beta,\sigma_*}(Y,X) = \mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X) \,|\, Y,X]$ sets this quantity to zero. It follows that $\widehat{\delta}^{\mathsf{P}}$ *minimizes the first-order contribution to worst-case bias.*

Global bound

• Our second result is that, for any fixed $\epsilon \geq 0$, the bias $b_\epsilon(h^{\mathsf{P}}_{\beta,\sigma_*})$ of the posterior estimator satisfies the *global bound*

$$b_\epsilon(h^{\mathsf{P}}_{\beta,\sigma_*}) \leq 2 \inf_h b_\epsilon(h).$$

• Hence, although $\widehat{\delta}^{\mathsf{P}}$ is not necessarily optimal in terms of bias under global misspecification, its worst-case bias is never worse than twice that of the best possible estimator.

• The proof relies on very different arguments than the proof in the small-ϵ case.

• Moreover, we show in a simple binary choice example that the factor 2 cannot be improved upon in general.

Conclusion ("Posterior Average Effects")

• Although posterior averaging is natural from a Bayesian perspective, its frequentist justification is unclear.

• We establish a local bias optimality result and a bias bound for posterior estimators of average effects. These results provide a rationale for the use of posterior estimators, beyond the settings to which they have been applied so far.