Minimizing sensitivity to model misspecification

STÉPHANE BONHOMME
Kenneth C. Griffin Department of Economics, University of Chicago

MARTIN WEIDNER
Department of Economics, University of Oxford and Institute for Fiscal Studies

We propose a framework for estimation and inference when the model may be misspecified. We rely on a local asymptotic approach where the degree of misspecification is indexed by the sample size. We construct estimators whose mean squared error is minimax in a neighborhood of the reference model, based on one-step adjustments. In addition, we provide confidence intervals that contain the true parameter under local misspecification. As a tool to interpret the degree of misspecification, we map it to the local power of a specification test of the reference model. Our approach allows for systematic sensitivity analysis when the parameter of interest may be partially or irregularly identified. As illustrations, we study three applications: an empirical analysis of the impact of conditional cash transfers in Mexico where misspecification stems from the presence of stigma effects of the program, a cross-sectional binary choice model where the error distribution is misspecified, and a dynamic panel data binary choice model where the number of time periods is small and the distribution of individual effects is misspecified.

KEYWORDS. Model misspecification, robustness, sensitivity analysis.

JEL CLASSIFICATION. C13, C23.

1. INTRODUCTION

Although economic models are intended as plausible approximations to a complex economic reality, econometric inference often relies on the model being an exact description of the population environment. To account for the possibility that their models are misspecified, economists have developed a number of approaches such as specifica-

tion tests, semi-parametric and nonparametric methods, and more recently bounds approaches. Implementing those approaches typically requires estimating a more general model than the original specification, possibly involving nonparametric and partially identified components.

In this paper, we consider a different approach, which consists in quantifying how model misspecification affects the parameter of interest, and in modifying the estimate in order to minimize the impact of misspecification. The goal of the analysis is twofold. First, we provide simple adjustments, which do not require reestimating the model, and provide guarantees on performance when the model is misspecified. Second, we construct confidence intervals that account for model misspecification error in addition to sampling uncertainty.

In our approach, we consider deviations from a reference specification of the model, in a particular class. The reference model is parametric and fully specified given covariates. It may, for example, correspond to the empirical specification of a structural economic model. We do not assume that the reference model is correctly specified, and allow for *local* deviations from it within a larger class of models. Relative to other approaches, a local analysis presents important advantages in terms of tractability.

We construct minimax estimators which minimize worst-case mean squared error (MSE) in a given neighborhood of the reference model. The worst case is influenced by the directions of model misspecification, which matter most for the parameter of interest. We focus in particular on two types of neighborhoods, for two leading classes of applications: Euclidean neighborhoods, in settings where the larger class of models containing the reference specification is parametric, and Kullback–Leibler neighborhoods, in semiparametric mixture models where misspecification of functional forms is measured by the Kullback–Leibler divergence between density functions.

The framework we propose is inspired by Hansen and Sargent's (2001, 2008) work on robust decision making under uncertainty and ambiguity. As in their work, optimal decisions depend on the size of the neighborhood around the reference model. In this paper, we do not attempt to provide a data-driven choice for the neighborhood size. Instead, we take the size as given and derive formulas for optimal estimation in neighborhoods of a given size. We discuss how to interpret the magnitude of the neighborhood size in various parametric and semiparametric examples. In addition, we show that the neighborhood size can be mapped to the local power—in certain directions—of a likelihood-ratio test of correct specification of the reference model.

Our approach delivers a class of estimators that can be used for systematic sensitivity analysis. In addition, we show how to construct confidence intervals which asymptotically contain the population parameter of interest with prespecified probability, both under correct specification and local misspecification. We show that acknowledging misspecification leads to easy-to-compute enlargements of conventional confidence intervals. Such confidence intervals are "honest," in the sense that they account for the bias of the estimator (e.g., Donoho (1994), Armstrong and Kolesár (2020)).

Our local approach leads to tractable expressions for worst-case bias and MSE, as well as for minimum-MSE estimators in a given neighborhood of the reference model. A minimum-MSE estimator takes the form of a one-step adjustment of the estimator

based on the reference model by a term, which reflects the impact of model misspecification, in addition to a more standard term, which adjusts the estimate in the direction of the efficient estimator based on the reference model. Implementing the optimal estimator only requires computing the score and Hessian of a larger model, evaluated at the reference model. The large model never needs to be estimated. This feature of our approach is reminiscent of the logic of Lagrange Multiplier (LM) testing.

We illustrate our approach using three examples. We first study the evaluation of the PROGRESA program in Mexico, which provides income transfers to households subject to the condition that the child attends school. Todd and Wolpin (2006) estimate a structural model of education choice on villages that were initially randomized out. They compare the predictions of the structural model with the estimated experimental impact. As emphasized by Todd and Wolpin (2008) and Attanasio et al. (2012), the ability to predict the effects of the program based solely on control villages imposes restrictions on the economic model. Within a simple static model of education choice, we assess the sensitivity of counterfactual predictions to a form of misspecification under which program participation may have a direct "stigma" effect on the marginal utility of schooling (Wolpin (2013)).

We next study the impact of misspecification of the error distribution in a cross-sectional binary choice model. Our aim is to estimate the outcome probabilities under different values of the covariates. While point-identification can be achieved under independence and sufficiently rich support of covariates (Manski (1988)), the quantities of interest are partially identified in our setting. Relying on a normal (probit) reference model, we show how our estimators and confidence intervals can be used for sensitivity analysis, when the researcher is concerned about misspecification of the normal distribution.

Our third and last example is a dynamic binary choice model in short panel data. We assume that time-varying errors are *i.i.d.* normal, but leave the distribution of individual heterogeneity given initial conditions unrestricted. In this setting also, common parameters and average effects often fail to be point-identified (Chamberlain (2010), Honoré and Tamer (2006), and Chernozhukov et al. (2013)), thus motivating a sensitivity analysis approach. We show that minimizing worst-case MSE in such panel data settings leads to a Tikhonov-regularized estimator, where the penalization reflects the degree of misspecification allowed for. In simulations, we illustrate that our estimator can provide substantial bias and MSE reduction relative to commonly used estimators.

Related work and outline   As in the literature on robust statistics (Huber (1964), Huber and Ronchetti (2009), Hampel et al. (1986), and especially Rieder (1994)), we rely on a minimax approach and aim to minimize the worst-case impact of misspecification in a neighborhood of a model. A difference with this work is that we focus on misspecification of specific aspects of a model, by considering parametric or semiparametric classes of models around the reference specification. By contrast, the robust statistics literature has mostly focused on fully nonparametric classes, motivated by data contamination issues.

A related literature studies orthogonalization and locally robust moment functions; see Neyman (1959), Newey (1994), Chernozhukov et al. (2018), Chernozhukov et al.

(2020), and also Fraser (1964). Here, we account for both bias and variance, weighting them by the size of the neighborhood around the reference model. In addition, our approach does not require the larger model to be point-identified. Our analysis also connects to Bayesian robustness (e.g., Berger and Berliner (1986), Gustafson (2000), Vidakovic (2000), Mueller (2012)), although our minimum-MSE estimators and confidence intervals have a frequentist interpretation.

Also related are the literatures on statistical decision theory (e.g., Wald (1950), Chamberlain (2000), Watson and Holmes (2016), Hansen and Marinacci (2016), and especially Hansen and Sargent (2008)) and the literature on sensitivity analysis in statistics and economics (e.g., Rosenbaum and Rubin (1983), Leamer (1985), Imbens (2003), Altonji et al. (2005), Nevo and Rosen (2012), Oster (2019), Masten and Poirier (2020, 2021)). Our analysis of minimum-MSE estimation and sensitivity in the OLS/IV example is related to Hahn and Hausman (2005) and Angrist et al. (2017). Our approach based on local misspecification has a number of precedents, such as Newey (1985), Conley et al. (2012), Guggenberger (2012), Bugni et al. (2012), Kitamura et al. (2013), and Bugni and Ura (2019). Also related is Claeskens and Hjort's (2003) work on the focused information criterion.

Recent papers rely on a local approach to misspecification to provide tools for sensitivity analysis. Andrews et al. (2017) propose a measure of sensitivity of parameter estimates to the moments used in estimation. Andrews et al. (2020) introduce a measure of informativeness of descriptive statistics in the estimation of structural models; see also Mukhin (2018). Our goal is different, in that we aim to provide a framework for estimation and inference in the presence of misspecification. Armstrong and Kolesár (2021) study models defined by over-identified systems of moment conditions that are approximately satisfied at true values, up to an additive term that vanishes asymptotically, and derive results for optimal estimation and inference. In this paper, we seek to ensure robustness to misspecification of a reference model within a larger class of models.

Our focus on specific forms of model misspecification also relates to recent approaches to estimate partially identified models (Chen et al. (2011), Norets and Tang (2014), Schennach (2014), and Giacomini and Kitagawa (2021)). Christensen and Connault (2019) consider structural models defined by equilibrium conditions, and develop inference methods on the identified set of counterfactual predictions subject to restrictions on the distance between the true model and a reference specification. Our local approach is complementary to these methods. It allows tractability in complex models, such as structural economic models, since implementation does not require estimating a larger model. In our framework, we view the parametric reference model as a useful benchmark, although its predictions need to be modified in order to minimize the impact of misspecification. This aspect relates our paper to shrinkage methods (e.g., Hansen (2016, 2017), Fessler and Kasy (2019), and Maasoumi (1978)), with the difference that here we are interested in a single parameter.

The plan of the paper is as follows. In Section 2, we describe our framework and derive the main results. In Section 3, we apply our framework to parametric and semiparametric mixture models. In Section 4, we discuss how to use our approach for sensitivity analysis, with a focus on the interpretation of neighborhood size. In Sections 5 and 6, we

present our illustrations. Finally, we conclude in Section 7. The Online Supplementary Material (Bonhomme and Weidner (2022)) contains an Appendix and codes for replication.

2. Framework of analysis

In this section, we describe the main elements of our approach in a general setting. In the next section, we will specialize the analysis to a locally quadratic setting, which includes both parametric misspecification and semiparametric misspecification of distributional functional forms.

2.1 *Setup*

We observe a random sample $(Y_i : i = 1, \ldots, n)$ from a density $f_{\beta,\pi}(y)$ (with respect to a continuous or discrete measure), where $\beta \in \mathcal{B}$ is a finite-dimensional parameter, and $\pi \in \Pi$ is a finite- or infinite-dimensional parameter. Throughout the paper, the parameter of interest is $\delta_{\beta,\pi}$, a scalar function or functional of β and π. We assume that $\delta_{\beta,\pi}$ and $f_{\beta,\pi}$ are known, smooth functions of β and π. Examples of functionals of interest in economic applications include counterfactual policy effects in structural models, and average effects in panel data settings. The true parameter values β_0 and π_0 that generate the observed data $Y_1, \ldots, Y_n$ are unknown to the researcher. Our goal is to estimate δ_{β_0,π_0} and construct confidence intervals for it. We abstract from covariates to simplify the presentation, but it is straightforward to extend our results to conditional models of the form $f_{\beta_0,\pi_0}(y|x)$; see Section 3.2.

Our starting point is that the researcher has chosen a reference model $\pi(\gamma)$, which parameterizes the unknown $\pi \in \Pi$ in terms of a finite-dimensional parameter $\gamma \in \mathcal{G}$. We say that the reference model is correctly specified if there exists a value $\gamma \in \mathcal{G}$ such that $\pi_0 = \pi(\gamma)$. Otherwise, we say that the model is misspecified. To measure misspecification, we rely on a distance measure d on Π, and we denote the maximal amount of misspecification as $\epsilon \geq 0$.

In our theory, we consider an asymptotic sequence where $\epsilon = \epsilon_n$ tends to zero as n tends to infinity, so the maximal amount of misspecification gets smaller as the sample size increases. The reason for focusing on ϵ tending to zero is tractability, as a small-ϵ analysis allows us to rely on linearization techniques and obtain simple, explicit expressions. Moreover, when estimating δ_{β_0,π_0}, the estimation bias due to misspecification (of order $\epsilon^{1/2}$) and the standard deviation (of order $n^{-1/2}$) are asymptotically comparable, so both play a role in the mean squared error. This local asymptotic approach has a number of precedents in the literature, notably Rieder (1994). Along the sequence, the true parameter $\pi_0 = \pi_{0,n}$ depends on n, and we assume that, for a fixed parameter γ_*, $d(\pi_{0,n}, \pi(\gamma_*)) \leq \epsilon_n$ for all n. This implies that $\lim_{n \to \infty} d(\pi_{0,n}, \pi(\gamma_*)) = 0$; that is, $\pi_{0,n}$ converges to $\pi(\gamma_*)$ as n tends to infinity. Hereafter, we drop the indices n and do not make the sample size dependence of ϵ and π_0 explicit. For example, we simply write $d(\pi_0, \pi(\gamma_*)) \leq \epsilon$.

Given the distance measure d, and some $\epsilon > 0$, we define an ϵ-neighborhood around $\pi(\gamma_*)$ as

$$\Gamma_\epsilon(\gamma_*) = \{\pi_0 \in \Pi : d(\pi_0, \pi(\gamma_*)) \leq \epsilon\}.$$

We assume that the true π_0 that generates the data satisfies $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Later we will assume that γ_* can be estimated consistently by some preliminary estimator $\widehat{\gamma}$. The distance measure d, the misspecification bound ϵ, and the preliminary estimator $\widehat{\gamma}$ are chosen by the researcher.[1]

EXAMPLES. As a first example, consider a parametric model defined by Euclidean parameters β and π, where $\pi = 0$ under the reference model. For example, π can represent the effect of an omitted control variable in a regression, or the degree of endogeneity of a regressor as in the example we analyze in Section 3.3. Suppose that the researcher is interested in the parameter $\delta_{\beta_0, \pi_0} = c'\beta_0$ for a known vector c, such as one component of β_0. In this case, we will take the weighted Euclidean (squared) distance $d(\pi_0, \pi) = \|\pi_0 - \pi\|_\Omega^2 = (\pi_0 - \pi)'\Omega(\pi_0 - \pi)$, for a positive-definite matrix Ω.

As a second example, consider a semiparametric mixture model whose likelihood depends on a finite-dimensional parameter vector β and a nonparametric density π of unobservables $A \in \mathcal{A}$, abstracting from conditioning covariates for simplicity. The joint density of (Y, A) is $g_{\beta_0}(y|a)\pi_0(a)$, for some known function g. Suppose that the researcher's goal is to estimate an average effect $\delta_{\beta_0, \pi_0} = \mathbb{E}_{\pi_0}\Delta(A, \beta_0)$, for a known function Δ. It is common to estimate the model by parameterizing the unknown density as $\pi(\gamma)$, where γ is finite-dimensional. We focus on situations where, although the researcher thinks of $\pi(\gamma)$ as a plausible approximation to the population distribution π_0, she is not willing to rule out that it may be misspecified. In this case, we use the Kullback–Leibler divergence to define semiparametric neighborhoods, and we take $d(\pi_0, \pi) = 2 \int_{\mathcal{A}} \log(\frac{\pi_0(a)}{\pi(a)})\pi_0(a)\,da$.

We focus on asymptotically linear estimators $\widehat{\delta} = \widehat{\delta}(Y_1, \ldots, Y_n)$ that admit a stochastic expansion of the form

$$\widehat{\delta} = \delta_{\beta_0, \pi(\gamma_*)} + \frac{1}{n}\sum_{i=1}^n h(Y_i, \beta_0, \gamma_*) + o_{P_0}(n^{-\frac{1}{2}} + \epsilon^{1/2}), \tag{1}$$

where this expansion holds uniformly for all $P_0 = P_{\beta_0, \pi_0}$ such that $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, in a sense that we will discuss below and make precise in Theorem 1. Along the sequence we consider, the product ϵn tends to a positive constant, so the remainder in (1) is $o_{P_0}(n^{-\frac{1}{2}})$.

---

[1]Instead of defining the ϵ-neighborhood of misspecified models around a fixed point $\pi(\gamma_*)$, one could alternatively consider all π_0 in the set $\Gamma_\epsilon = \cup_{\gamma \in \mathcal{G}}\Gamma_\epsilon(\gamma)$, which is the ϵ-neighborhood around the manifold of reference models $\pi(\gamma)$, $\gamma \in \mathcal{G}$. This alternative definition would avoid having to define γ_*, and we employed it in the first version of this paper to justify the same local approximation to the worst-case MSE optimal estimator derived below (see Section 2.3 in Bonhomme and Weidner (2018)). In the current presentation, we only introduce the ϵ-neighborhood $\Gamma_\epsilon(\gamma_*)$ around a fixed γ_*; however, this has no effect on the minimax misspecification adjustments that we derive. By fixing γ_*, this presentation also aligns with the way in which local minimax results are typically discussed in statistics (see, e.g., Theorem 8.11 in Van der Vaart (2007)).

Although asymptotic linearity is satisfied by many econometric estimators, it can fail in certain semi-parametric problems (e.g., Cattaneo et al., 2014) and in problems involving model selection or shrinkage (e.g., Liao (2013), Cheng and Liao (2015)), for example.

Equation (1) is a form of local regularity of the estimator $\widehat{\delta}$. Consider first the correctly specified case, where $\epsilon = 0$ and $\pi_0 = \pi(\gamma_*)$. In this case, $h(\cdot, \beta_0, \gamma_*)$ is the influence function of $\widehat{\delta}$. We assume that the following conditions are satisfied:

$$\mathbb{E}_{\beta_0, \pi(\gamma_*)} h(Y, \beta_0, \gamma_*) = 0, \tag{2}$$

and

$$\nabla_{\beta\gamma} \delta_{\beta_0, \pi(\gamma_*)} + \mathbb{E}_{\beta_0, \pi(\gamma_*)} \nabla_{\beta\gamma} h(Y, \beta_0, \gamma_*) = 0, \tag{3}$$

where $\mathbb{E}_{\beta, \pi}$ denotes the expectation under $f_{\beta, \pi}$, and $\nabla_{\beta\gamma}$ denotes the derivative with respect to the vector $(\beta', \gamma')'$. Both (2) and (3) are standard properties of influence functions of regular asymptotically linear estimators.

We will refer to (2) as *unbiasedness*, since it guarantees that $\widehat{\delta}$ is asymptotically unbiased for δ_{β_0, π_0} under correct specification of the reference model. We assume that unbiasedness holds at all possible values of β_0 and γ_*. Then, by differentiating (2) with respect to β_0 and γ_* and plugging the resulting equations into (3), we obtain

$$\mathbb{E}_{\beta_0, \pi(\gamma_*)} h(Y, \beta_0, \gamma_*) \nabla_{\beta\gamma} \log f_{\beta_0, \pi(\gamma_*)}(Y) = \nabla_{\beta\gamma} \delta_{\beta_0, \pi(\gamma_*)}. \tag{4}$$

Under unbiasedness, (3) and (4) are equivalent. We will later work with (4), since it only features $h(Y, \beta_0, \gamma_*)$ and not its gradient. Under suitable conditions, (4) is necessary and sufficient for the asymptotically linear estimator $\widehat{\delta}$ to be regular; see, e.g., Newey (1990). As an example, for m-estimators, (4) can be interpreted as the generalized information matrix equality. Asymptotic linearity and regularity are commonly imposed in the semi-parametric efficiency literature (Bickel et al. (1993)). These conditions rule out, for example, superefficient estimators such as Hodges' estimator. We will refer to (3), or alternatively (4), as *local robustness*, using a terminology introduced by Chernozhukov et al. (2020).[2]

Consider now the misspecified case, where $\epsilon > 0$. In this case, we strengthen the condition of asymptotic linearity, and require that $\widehat{\delta}$ be locally asymptotically linear; see, e.g., Klaassen (1987). Formally, under local, small-ϵ misspecification, we assume the stochastic expansion (1) continues to hold, but now uniformly for all $\pi_0 \in \Gamma_\epsilon(\gamma_*)$.[3] In the following, we focus on locally asymptotically linear estimators that satisfy (1), under the

---

[2]While in Chernozhukov et al. (2020) local robustness is imposed as a substantive restriction on more general moment functions, in our setting (3) and (4) are regularity conditions given unbiasedness.

[3]To see why (1) is a plausible way of imposing asymptotic linearity here, let $\phi(Y_i, \beta_0, \pi_0)$ be the influence function of $\widehat{\delta}$. Expanding as $n \to \infty$ and $\epsilon \to 0$, we have

$$\widehat{\delta} = \delta_{\beta_0, \pi(\gamma_*)} + \frac{1}{n} \sum_{i=1}^{n} \underbrace{\phi(Y_i, \beta_0, \pi(\gamma_*))}_{=h(Y_i, \beta_0, \gamma_*)} + [\pi_0 - \pi(\gamma_*)]' \underbrace{[\nabla_\pi \delta_{\beta_0, \pi(\gamma_*)} + \mathbb{E}_{\beta_0, \pi(\gamma_*)} \nabla_\pi \phi(Y, \beta_0, \pi(\gamma_*))]}_{=0}$$

$$+ o_{P_0}\left(n^{-\frac{1}{2}} + \epsilon^{\frac{1}{2}}\right).$$

conditions (2) and (4). Notice that, under local misspecification, the influence function $h(Y, \beta_0, \gamma_*)$ has no longer mean zero under P_0 in general.

Our goal in this paper is twofold. First, we will construct confidence intervals for the target parameter δ_{β_0, π_0} which are uniformly asymptotically valid on $\Gamma_\epsilon(\gamma_*)$. Second, an important goal of the analysis is to construct estimators $\widehat{\delta} = \widehat{\delta}(Y_1, \ldots, Y_n)$ that are asymptotically optimal in a minimax sense. For this purpose, we will show how to compute a function h such that the (trimmed) worst-case mean squared error (MSE) $\mathbb{E}_{\beta_0, \pi_0}[(\widehat{\delta} - \delta_{\beta_0, \pi_0})^2]$ in the ϵ-neighborhood $\Gamma_\epsilon(\gamma_*)$ of the reference model, among estimators of the form

$$\widehat{\delta}_{h, \widehat{\beta}, \widehat{\gamma}} = \delta_{\widehat{\beta}, \pi(\widehat{\gamma})} + \frac{1}{n} \sum_{i=1}^{n} h(Y_i, \widehat{\beta}, \widehat{\gamma}), \tag{5}$$

is minimized under our local asymptotic analysis. In fact, we will show how to compute estimators that minimize (trimmed) worst-case MSE among asymptotically linear estimators; see Theorem 1 below for a precise statement. Here, $\widehat{\beta}$ and $\widehat{\gamma}$ are preliminary estimators of β_0 and γ_* that are root-n consistent under correct specification. For example, $\widehat{\beta}$ and $\widehat{\gamma}$ may be maximum likelihood estimators (MLE) based on the reference model. It follows from (2) and (4) that, under regularity conditions on the preliminary estimators, the form of the minimum-MSE h function is not affected by the choice of $\widehat{\beta}$ and $\widehat{\gamma}$.

EXAMPLES (cont). In our parametric example, a natural estimator is the MLE of $c'\beta_0$ based on the reference specification, such as the OLS estimator under the assumption that $\pi = 0$; for example, that the coefficient of an omitted control variable is zero. In a correctly specified likelihood setting, such an estimator will be consistent and efficient. However, when the reference model is misspecified, it may be dominated in terms of bias or MSE by other regular estimators.

In our semiparametric mixture example, a commonly used ("random-effects") estimator of $\delta_{\beta_0, \pi_0} = \mathbb{E}_{\pi_0} \Delta(A, \beta_0)$ is obtained by replacing the population average by an integral with respect to the parametric distribution $\pi(\widehat{\gamma})$, where $\widehat{\gamma}$ is the MLE of γ. Another popular ("empirical Bayes") estimator is obtained by substituting an integral with respect to the posterior distribution of A based on $\pi(\widehat{\gamma})$. We will compare the finite-sample performance of these estimators to that of our minimum-MSE estimator in our panel data illustration in Section 6.

2.2 Heuristic derivation of the minimum-MSE estimator

In this subsection, we provide heuristic derivations for the worst-case bias and the minimum-MSE estimator. This will lead to the main expressions in equations (8) and (10) below. In the next subsection, we will provide regularity conditions under which these derivations are formally justified.

In this expansion, the term linear in $\pi_0 - \pi(\gamma_*)$ vanishes, whenever $\phi(Y, \beta_0, \pi_0)$ satisfies an influence function regularity condition analogous to (3), and the term quadratic in $\pi_0 - \pi(\gamma_*)$ gives a contribution $o_{P_0}(\epsilon^{\frac{1}{2}})$.

We assume that $\Gamma_\epsilon(\gamma_*)$ is a convex set. For any linear map $u : \Pi \to \mathbb{R}$, we define[4]

$$\|u\|_{\gamma_*,\epsilon} = \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \epsilon^{-\frac{1}{2}} u'\big(\pi_0 - \pi(\gamma_*)\big), \qquad \|u\|_{\gamma_*} = \lim_{\epsilon \to 0} \|u\|_{\gamma_*,\epsilon}. \tag{6}$$

We assume that the distance measure d is chosen such that $\|\cdot\|_{\gamma_*}$ is unique and well-defined, and that it is a norm, dual to a local approximation to $d(\pi_0, \pi(\gamma_*))$ for fixed $\pi(\gamma_*)$. Both our examples of distance measures—weighted Euclidean distance and Kullback–Leibler divergence—satisfy these assumptions.

We focus on estimators $\widehat{\delta}$ that satisfy (1) for a suitable h function for which (2) and (4) hold. Under appropriate regularity conditions, the worst-case bias of $\widehat{\delta}$ in the neighborhood $\Gamma_\epsilon(\gamma_*)$ can be expanded for small ϵ and large n as

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} |\mathbb{E}_{\beta_0, \pi_0} \widehat{\delta} - \delta_{\beta_0, \pi_0}| = b_\epsilon(h, \beta_0, \gamma_*) + o\big(n^{-\frac{1}{2}} + \epsilon^{\frac{1}{2}}\big), \tag{7}$$

where

$$b_\epsilon(h, \beta_0, \gamma_*) = \epsilon^{\frac{1}{2}} \big\| \nabla_\pi \delta_{\beta_0, \pi(\gamma_*)} - \mathbb{E}_{\beta_0, \pi(\gamma_*)} h(Y, \beta_0, \gamma_*) \nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(Y) \big\|_{\gamma_*}, \tag{8}$$

for $\|\cdot\|_{\gamma_*}$ the dual norm defined in (6).[5]

Then the worst-case MSE in $\Gamma_\epsilon(\gamma_*)$ can be expanded as follows, again under appropriate regularity conditions (see Lemma A2 in the Appendix):

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[(\widehat{\delta} - \delta_{\beta_0, \pi_0})^2\big] = b_\epsilon(h, \beta_0, \gamma_*)^2 + \frac{\mathrm{Var}_{\beta_0, \pi(\gamma_*)}\big(h(Y, \beta_0, \gamma_*)\big)}{n}$$

$$+ o\big(n^{-1} + \epsilon\big). \tag{9}$$

We therefore define the minimum-MSE function $h_\epsilon^{\mathrm{MMSE}}(y, \beta_0, \gamma_*)$ as

$$h_\epsilon^{\mathrm{MMSE}}(\cdot, \beta_0, \gamma_*) = \operatorname*{argmin}_{h(\cdot, \beta_0, \gamma_*)} \Big\{ \epsilon \big\| \nabla_\pi \delta_{\beta_0, \pi(\gamma_*)} - \mathbb{E}_{\beta_0, \pi(\gamma_*)} h(Y, \beta_0, \gamma_*) \nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(Y) \big\|_{\gamma_*}^2$$

$$+ \frac{\mathrm{Var}_{\beta_0, \pi(\gamma_*)}\big(h(Y, \beta_0, \gamma_*)\big)}{n} \Big\} \quad \text{subject to (2) and (4)}. \tag{10}$$

Finally, let $\widehat{\beta}$ and $\widehat{\gamma}$ be preliminary estimators that are consistent for β_0 and γ_* under the reference model $f_{\beta_0, \pi(\gamma_*)}$. Then the minimum-MSE estimator of δ_{β_0, π_0} is given by

$$\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \delta_{\widehat{\beta}, \pi(\widehat{\gamma})} + \frac{1}{n} \sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i, \widehat{\beta}, \widehat{\gamma}). \tag{11}$$

This estimator minimizes an asymptotic approximation to the worst-case MSE in $\Gamma_\epsilon(\gamma_*)$. Using a small-$\epsilon$ approximation is crucial for analytic tractability, since the variance term in (9) only needs to be calculated under the reference model, and the optimization problem (10) is convex. In practice, (10) only needs to be solved at $\widehat{\beta}$ and $\widehat{\gamma}$.

---

[4]The definition in (6) is stated for finite-dimensional π. This definition can be generalized to the case where π is infinite-dimensional; see the Appendix and Section 3.

[5]When π is infinite-dimensional ∇_π denotes a Gâteaux derivative.

In addition, as we already pointed out, the form of the minimum-MSE estimator is not affected by the choice of the preliminary estimators $\widehat{\beta}$ and $\widehat{\gamma}$.

The constraints on $h_\epsilon^{\mathrm{MMSE}}(\cdot, \beta_0, \gamma_*)$ imposed in (10) are the unbiasedness condition (2) and the local robustness condition (4). As we discussed above, given unbiasedness, local robustness is a regularity condition, and unbiasedness is a substantive condition that implies that our estimator $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is only optimal within the class of estimators that are asymptotically unbiased for δ_{β_0, π_0} under the reference model.

Special cases   To provide intuition about the minimum-MSE function $h_\epsilon^{\mathrm{MMSE}}$, let us define two Hessian matrices $H_{\beta\gamma}$, of size $\dim \beta + \dim \gamma$, and $\mathcal{H}_{\beta\pi}$, of size $\dim \beta + \dim \pi$, as[6]

$$H_{\beta\gamma} = \mathbb{E}_{\beta_0, \pi(\gamma_*)}\big[\nabla_{\beta\gamma} \log f_{\beta_0, \pi(\gamma_*)}(Y)\big]\big[\nabla_{\beta\gamma} \log f_{\beta_0, \pi(\gamma_*)}(Y)\big]',$$

$$\mathcal{H}_{\beta\pi} = \mathbb{E}_{\beta_0, \pi(\gamma_*)}\big[\nabla_{\beta\pi} \log f_{\beta_0, \pi(\gamma_*)}(Y)\big]\big[\nabla_{\beta\pi} \log f_{\beta_0, \pi(\gamma_*)}(Y)\big]'.$$

Throughout our analysis, we assume that $H_{\beta\gamma}$ is invertible. This requires that the Hessian matrix of the parametric reference model be nonsingular, thus requiring that β_0 and γ_* be identified under the reference model. When $\epsilon = 0$, we find that

$$h_0^{\mathrm{MMSE}}(y, \beta_0, \gamma_*) = \nabla_{\beta\gamma} \log f_{\beta_0, \pi(\gamma_*)}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma} \delta_{\beta_0, \pi(\gamma_*)}. \tag{12}$$

Thus, under the assumption that the parametric reference model is correctly specified, $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is simply the one-step approximation to the MLE for δ_{β_0, π_0} that maximizes the likelihood with respect to the "small" parameter $(\beta', \gamma')'$. This "one-step efficient" adjustment is purely based on efficiency considerations. Such one-step approximations are classical estimators in statistics (e.g., Bickel et al. (1993)).

Another interesting special case of the minimum-MSE h function arises in the limit $\epsilon \to \infty$, when the matrix or operator $\mathcal{H}_{\beta\pi}$ is invertible. Note that invertibility of $\mathcal{H}_{\beta\pi}$, which may fail when π_0 is not identified, is not needed in our analysis and we only use it to analyze this limiting case. We then have that

$$\lim_{\epsilon \to \infty} h_\epsilon^{\mathrm{MMSE}}(y, \beta_0, \gamma_*) = \big[\nabla_{\beta\pi} \log f_{\beta_0, \pi(\gamma_*)}(y)\big]' \mathcal{H}_{\beta\pi}^{-1} \nabla_{\beta\pi} \delta_{\beta_0, \pi(\gamma_*)}. \tag{13}$$

In this limit, $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is simply the one-step approximation to the MLE for δ_{β_0, π_0} that maximizes the likelihood with respect to the "large" parameter $(\beta', \pi')'$. For any ϵ, the estimator $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is a nonlinear interpolation between the one-step MLE approximation to the parametric reference model and the one-step MLE approximation to the large model. We obtain one-step approximations in our approach, since (10) is only a local approximation to the full MSE-minimization problem.

However, an estimator based on (13) may be ill-behaved in nonpoint-identified problems, or in problems where the identification of π_0 is irregular. By contrast, $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is always well-defined, since the variance of $h(Y, \beta_0, \gamma_*)$ acts as a sample size-dependent regularization. The form of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is thus based on both efficiency and robustness.

---

[6]The definition of $\mathcal{H}_{\beta\pi}$ generalizes to the infinite-dimensional π case; see the Appendix and Section 3.

In addition, note that while neither (12) nor (13) involve the particular choice of distance measure with respect to which neighborhoods are defined, for given $\epsilon > 0$, the minimum-MSE estimator will depend on the chosen distance measure.

Lastly, it is common in applications with covariates to model the conditional distribution of outcomes Y given covariates X as $f_{\beta_0, \pi_0}(y|x)$, while leaving the marginal distribution of X, $f_X(x)$, unspecified. Our approach can easily be adapted to deal with such conditional models, as we will describe in Section 3.2 in a locally quadratic setting.

2.3 *Properties of the minimum-MSE estimator*

In this subsection, we provide a formal characterization of the minimum-MSE estimator by showing that it achieves minimum worst-case MSE in a class of regular asymptotically linear estimators, as n tends to infinity and ϵn tends to a constant. All sequences can thus be equivalently indexed by ϵ or n; for example, h_ϵ in the following theorem could equivalently be indexed by n. Moreover, under the stated assumptions, the heuristic derivations of the previous subsection are formally justified. All proofs are in the Appendix.

THEOREM 1. *Let $n \to \infty$ and $\epsilon \to 0$ such that $\epsilon n \to c$, for some constant $c \in (0, \infty)$. Let Assumptions A1 and A2 in the Appendix hold, let $\beta_0 \in \mathcal{B}$ and $\gamma_* \in \mathcal{G}$, and let $\widehat{\delta}_\epsilon = \widehat{\delta}_\epsilon(Y_1, \ldots, Y_n)$ be a sequence of estimators with an influence function expansion of the form*

$$\widehat{\delta}_\epsilon = \delta_{\beta_0, \pi(\gamma_*)} + \frac{1}{n} \sum_{i=1}^n h_\epsilon(Y_i, \beta_0, \gamma_*) + n^{-1/2} R_n, \tag{14}$$

where R_n is a sequence of random variables with

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} P_{\beta_0, \pi_0}\big(|R_n| > \log(n)\big) = o(1), \qquad \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[R_n^2 \mathbb{1}\big(|R_n| \leq 2\log(n)\big)\big] = o(1),$$

and $h_\epsilon(\cdot, \beta_0, \gamma_*)$ is a sequence of influence functions that satisfy the constraints (2) and (4), as well as $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}|h_\epsilon(Y, \beta_0, \gamma_*)|^\kappa = O(1)$, for some $\kappa > 2$. We then have, for any sequence $m_n > 0$ with $m_n \to 0$ and $m_n n^{1/2}[\log(n)]^{-1} \to \infty$, that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \delta_{\beta_0, \pi_0}\big)^2 \mathbb{1}\big(\big|\widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \delta_{\beta_0, \pi_0}\big| \leq m_n\big)\big]$$

$$\leq \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[(\widehat{\delta}_\epsilon - \delta_{\beta_0, \pi_0})^2 \mathbb{1}\big(|\widehat{\delta}_\epsilon - \delta_{\beta_0, \pi_0}| \leq m_n\big)\big] + o\Big(\frac{1}{n}\Big). \tag{15}$$

We establish Theorem 1 in a joint asymptotic sequence where ϵ tends to zero as n tends to infinity and ϵn tends to a finite positive constant. Under this sequence, the leading term in the worst-case MSE is of order ϵ (squared bias), or equivalently of order $1/n$ (variance). The theorem considers a trimmed MSE to allow for the possibility that the estimators for δ_{β_0, π_0} do not have moments. The trimming cutoff m_n shrinks to zero at a rate slower than $n^{-1/2}$ (or equivalently $\epsilon^{1/2}$), so that for estimators without heavy

tails the leading-order bias and standard deviation should not be affected by the trimming. The theorem states that the leading-order worst-case trimmed MSE achieved by our minimum-MSE estimator $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is at least as good as the one achieved by any other sequence of estimators satisfying our regularity conditions. All the assumptions on $\widehat{\delta}_\epsilon$ and $h_\epsilon(\cdot, \beta, \gamma)$ that we require for this result are listed in the statement of the theorem. Note that in Theorem 1 we assume that (14) holds, subject to (2) and (4). On might conjecture that the result holds absent these conditions. However, our current proof crucially relies on them.[7] We will provide explicit expressions for $h_\epsilon^{\mathrm{MMSE}}$ in various models in the next section.

2.4 *Confidence intervals*

In addition to point estimates, our framework allows us to compute confidence intervals that contain δ_{β_0, π_0} with prespecified probability under our local asymptotic analysis. To see this, let $\widehat{\delta}$ be an estimator satisfying (1), (2), and (4). For a given confidence level $\alpha \in (0, 1)$, let us define the following interval:

$$CI_\epsilon(1 - \alpha, \widehat{\delta}) = \left[\widehat{\delta} \pm \left(b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) + \frac{\widehat{\sigma}_h}{\sqrt{n}} c_{1-\alpha/2} \right) \right], \tag{16}$$

where $b_\epsilon(\cdot)$ is given by (8), $\widehat{\sigma}_h^2$ is the sample variance of $h(Y_1, \widehat{\beta}, \widehat{\gamma}), \ldots, h(Y_n, \widehat{\beta}, \widehat{\gamma})$, and $c_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ is the $(1 - \alpha/2)$-standard normal quantile. Under suitable regularity, conditions, the interval $CI_\epsilon(1 - \alpha, \widehat{\delta})$ contains δ_{β_0, π_0} with probability at least $1 - \alpha$ as n tends to infinity and ϵn tends to a constant, both under correct specification and under local misspecification of the reference model. Formally, we have the following result.

THEOREM 2. *Let $n \to \infty$ and $\epsilon \to 0$ such that $\epsilon n \to c$, for some constant $c \in (0, \infty)$. Let Assumptions* A1 *and* A3 *in the Appendix hold, and also assume that the influence function h of $\widehat{\delta}$ satisfies $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} h^2(Y, \beta_0, \gamma_*) = O(1)$. Then we have*

$$\inf_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \Pr_{\beta_0, \pi_0} \left[\delta_{\beta_0, \pi_0} \in CI_\epsilon(1 - \alpha, \widehat{\delta}) \right] \geq 1 - \alpha + o(1). \tag{17}$$

Such "fixed-length" confidence intervals, which take into account both misspecification bias and sampling uncertainty, have been studied in different contexts (e.g., Donoho (1994), Armstrong and Kolesár (2020, 2021)).[8]

---

[7]Condition (14) is a form of local regularity of the sequence of estimators $\widehat{\delta}_\epsilon$. The additional regularity conditions in Assumptions A1 and A2 are smoothness conditions on $f_{\beta_0, \pi_0}(y)$, δ_{β_0, π_0}, $\pi(\gamma)$, and $d(\pi_0, \pi(\gamma))$ as functions of β_0, π_0, and γ, and an appropriate rate condition on the preliminary estimators $\widehat{\beta}$ and $\widehat{\gamma}$. In particular, in Assumption A2, we require the preliminary estimators $\widehat{\beta}$ and $\widehat{\gamma}$ to have moments of order larger than two. This may require modifying the preliminary estimators to ensure that they have finite moments, as in, for example, Hausman et al. (2011), who focus on GMM estimators.

[8]A variation suggested by these authors, which reduces the length of the interval, is to compute the interval as $\widehat{\delta} \pm b_\epsilon(h, \widehat{\beta}, \widehat{\gamma})$ times the $(1 - \alpha)$-quantile of $|\mathcal{N}(1, \widehat{\sigma}_h^2/(nb_\epsilon(h, \widehat{\beta}, \widehat{\gamma})^2))|$.

3. Locally quadratic case: Applications to parametric models and semiparametric mixture models

In this section, we first derive explicit expressions for minimum-MSE estimators in a class of models that have a locally quadratic structure; that is, where the dual norm $\|\cdot\|_{\gamma_*}$ can be associated with an inner product. We then apply these results to parametric models and semiparametric mixture models. Our motivation for focusing on these settings is that they allow us to develop practical implementation methods. We have not explored implementation in other models with infinite-dimensional π parameters.

Formally, let us define the tangent space $\overline{\mathcal{T}}$ of the parameter space Π at $\pi(\gamma^*)$, where for simplicity we ignore the dependence on γ^* in the notation.[9] Let us then define the cotangent space $\mathcal{T}$ as the set of linear maps $u : \overline{\mathcal{T}} \to \mathbb{R}$. Throughout this section, we assume that $\mathcal{T}$ is a Hilbert space equipped with the norm $\|\cdot\|_{\gamma_*}$. This locally quadratic structure characterizes our two leading examples of parametric and semiparametric mixture models. In such cases, the tangent space (at an interior $\pi(\gamma^*)$) is simply Π, and the cotangent space is the set of linear maps $u : \Pi \to \mathbb{R}$.

3.1 *Characterization of the minimum-MSE estimator*

Consider the case where the square of the local dual norm defined in (6) can be written as $\|u\|_{\gamma_*}^2 = u^\top u$, where $u^\top w$ represents some inner product of elements u and w of the cotangent space $\mathcal{T}$ of Π at $\pi(\gamma_*)$. For conciseness, from now on we will remove the subscripts β_0, γ_*, and $\pi(\gamma_*)$ throughout, unless there is a risk of confusion. In particular, unless otherwise noted, all expectations will be evaluated under the reference model. Here, π can be finite-dimensional as in parametric models (which we analyze in Section 3.3), or infinite-dimensional as in semiparametric mixture models where π is a density (studied in Section 3.4).

Let us start by introducing some notation. Let $s_{\beta\gamma}(y) = \nabla_{\beta\gamma} \log f(y)$ and $s_\pi(y) = \nabla_\pi \log f(y)$ denote the components of the score. We define the Hessian operators $H_\pi : \mathcal{T} \to \mathcal{T}$, $H_{\pi,\beta\gamma} : \mathbb{R}^{\dim \beta + \dim \gamma} \to \mathcal{T}$, and $H_{\beta\gamma,\pi} : \mathcal{T} \to \mathbb{R}^{\dim \beta + \dim \gamma}$ by[10]

$$H_\pi = \mathbb{E} s_\pi(Y) s_\pi(Y)^\top, \qquad H_{\pi,\beta\gamma} = \mathbb{E} s_\pi(Y) s_{\beta\gamma}(Y)', \qquad H_{\beta\gamma,\pi} = \mathbb{E} s_{\beta\gamma}(Y) s_\pi(Y)^\top.$$

In addition, we define the following projected versions of the gradient $\widetilde{\nabla}_\pi = \nabla_\pi - H_{\pi,\beta\gamma} H_{\beta\gamma}^{-1} \nabla_{\beta\gamma}$, score $\widetilde{s}_\pi(y) = s_\pi(y) - H_{\pi,\beta\gamma} H_{\beta\gamma}^{-1} s_{\beta\gamma}(y)$, and Hessian $\widetilde{H}_\pi = H_\pi - H_{\pi,\beta\gamma} \times H_{\beta\gamma}^{-1} H_{\beta\gamma,\pi}$.

The next lemma characterizes the minimum-MSE h function in the locally quadratic case.

---

[9] The tangent space at $\pi(\gamma^*)$ includes all directions in which one can pass tangentially through $\pi(\gamma^*)$. See Barden and Thomas (2003) for a formal exposition.

[10] Formally, $u^\top$ is an element of the tangent space of Π at $\pi(\gamma_*)$; that is, $u \mapsto u^\top$ represents a linear mapping from the cotangent space $\mathcal{T}$ to the tangent space $\overline{\mathcal{T}}$.

Lemma 1. *For the locally quadratic case of this section, the three following equivalent characterizations of h_ϵ^{MMSE} defined in (10) hold:*

$$h_\epsilon^{\text{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma} \delta + (\epsilon n) \widetilde{s}_\pi(y)^\top \big(\widetilde{\nabla}_\pi \delta - \mathbb{E}\big[h_\epsilon^{\text{MMSE}}(Y)\widetilde{s}_\pi(Y)\big]\big) \tag{18}$$

$$= s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma} \delta + (\epsilon n) \widetilde{s}_\pi(y)^\top \big(\nabla_\pi \delta - \mathbb{E}\big[h_\epsilon^{\text{MMSE}}(Y) s_\pi(Y)\big]\big) \tag{19}$$

$$= s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma} \delta + \widetilde{s}_\pi(y)^\top \big[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I}\big]^{-1} \widetilde{\nabla}_\pi \delta, \tag{20}$$

where $\mathbb{I}$ denotes the identity map on $\mathcal{T}$.

3.2 *Covariates*

So far in our presentation, we have abstracted from covariates. We now consider the case where in addition to the outcomes Y_i we observe a vector of covariates X_i. We assume that (Y_i, X_i) are randomly drawn from a conditional distribution of Y_i given X_i given by the model $f_{\beta_0,\pi_0}(y|x)$, and an unrestricted marginal distribution f_X of X_i. Our parameter of interest is $\delta_{\beta_0,\pi_0,f_X} = \mathbb{E}_{f_X} \delta_{\beta_0,\pi_0}(X)$, where $\mathbb{E}_{f_X}$ denotes an expectation over f_X. We consider estimators of the form

$$\widehat{\delta}_h = \frac{1}{n}\sum_{i=1}^n \delta_{\widehat{\beta},\pi(\widehat{\gamma})}(X_i) + \frac{1}{n}\sum_{i=1}^n h(Y_i, X_i, \widehat{\beta}, \widehat{\gamma}, \widehat{f}_X),$$

where $\widehat{\beta}$ and $\widehat{\gamma}$ are preliminary estimates whose probability limits are β_0 and γ_*, and $\widehat{f}_X$ is the empirical distribution of X_i in the sample. While f_X is unknown and infinite-dimensional, it only enters into our object of interest (and the expression of h_ϵ^{MMSE} below) as an expectation, and the corresponding sample average is still estimated at the $\sqrt{n}$-rate. We have the following characterization of the minimum-MSE influence function.

Lemma 2. *Consider the locally quadratic case of this section. Let $H_{\beta\gamma}(x)$ and $H_{\pi,\beta\gamma}(x)$ be conditional counterparts to $H_{\beta\gamma}$ and $H_{\pi,\beta\gamma}$, and likewise let $s_{\beta\gamma}(y|x)$, $s_\pi(y|x)$ and $\widetilde{s}_\pi(y|x) = s_\pi(y|x) - [\mathbb{E}_{f_X} H_{\pi,\beta\gamma}(X)][\mathbb{E}_{f_X} H_{\beta\gamma}(X)]^{-1} s_{\beta\gamma}(y|x)$ denote (projected) scores in the conditional model. We have*

$$\begin{aligned} h_\epsilon^{\text{MMSE}}(y, x) &= \delta(x) - \mathbb{E}_{f_X} \delta(X) + s_{\beta\gamma}(y|x)'\big[\mathbb{E}_{f_X} H_{\beta\gamma}(X)\big]^{-1}\mathbb{E}_{f_X} \nabla_{\beta\gamma} \delta(X) \\ &\quad + (\epsilon n)\widetilde{s}_\pi(y|x)^\top\big\{\mathbb{E}_{f_X}\widetilde{\nabla}_\pi \delta(X) - \mathbb{E}_{f_X}\mathbb{E}\big[h_\epsilon^{\text{MMSE}}(Y, X)\widetilde{s}_\pi(Y|X)\big]\big\}, \end{aligned} \tag{21}$$

with analogous counterparts to (19) and (20).

A first difference between (18) and (21) is that various expectations over f_X occur here, which we will replace by sample averages when calculating the estimator $\widehat{\delta}_\epsilon^{\text{MMSE}}$ in practice. A second difference comes from the term $\delta(x) - \mathbb{E}_{f_X} \delta(X)$. However, this term does not contribute to $\widehat{\delta}_\epsilon^{\text{MMSE}}$, since its sample average is zero once we replace f_X by the empirical distribution $\widehat{f}_X$.[11]

---

[11]The term $\delta(x) - \mathbb{E}_{f_X} \delta(X)$ ensures that h_ϵ^{MMSE} is locally robust with respect to f_X in the sense of (4).

3.3 *Parametric models*

A simple locally quadratic example is a parametric model where π is finite-dimensional, and the distance measure over π is based on a weighted Euclidean metric $\|\cdot\|_\Omega$ for a positive definite weight matrix Ω. Here, we treat Ω and the neighborhood size ϵ as known. We will discuss the choice of Ω, and the interpretation of ϵ, in Section 4.

The small-ϵ approximation to the bias of $\widehat{\delta}$ is given by (8), with $\|\cdot\|_{\gamma_*} = \|\cdot\|_{\Omega^{-1}}$, where Ω^{-1} is the inverse of Ω. In this case, for vectors $u, w \in \mathbb{R}^{\dim \pi}$ we have $u^\top w = u'\Omega^{-1}w$. Let

$$\mathbb{H}_\pi = \mathbb{E}\big[s_\pi(Y)s_\pi(Y)'\big],$$

$$\widetilde{\mathbb{H}}_\pi = \mathbb{E}\big[s_\pi(Y)s_\pi(Y)'\big] - \mathbb{E}\big[s_\pi(Y)s_{\beta\gamma}(Y)'\big]H_{\beta\gamma}^{-1}\mathbb{E}\big[s_{\beta\gamma}(Y)s_\pi(Y)'\big],$$

be the usual parametric Hessian matrices. We have $H_\pi = \mathbb{H}_\pi\Omega^{-1}$, $H_{\pi,\beta\gamma} = \mathbb{E}[s_\pi(Y) \times s_{\beta\gamma}(Y)']$, $H_{\beta\gamma,\pi} = H'_{\pi,\beta\gamma}\Omega^{-1}$, $\widetilde{s}_\pi = s_\pi - H_{\pi,\beta\gamma}H_{\beta\gamma}^{-1}s_{\beta\gamma}$, and $\widetilde{H}_\pi = \widetilde{\mathbb{H}}_\pi\Omega^{-1}$. From (20) we then obtain the following.

COROLLARY 1 (Parametric models).

$$h_\epsilon^{\mathrm{MMSE}}(y) = s_{\beta\gamma}(y)'H_{\beta\gamma}^{-1}\nabla_{\beta\gamma}\delta + \widetilde{s}_\pi(y)'\Omega^{-1}\big[\widetilde{H}_\pi + (\epsilon n)^{-1}I\big]^{-1}\widetilde{\nabla}_\pi\delta,$$

where I is the identity matrix of size $\dim \pi$.

In addition to the "one-step efficient" adjustment $h_0^{\mathrm{MMSE}} = s'_{\beta\gamma}H_{\beta\gamma}^{-1}\nabla_{\beta\gamma}\delta$, the minimum-MSE function $h_\epsilon^{\mathrm{MMSE}}$ in Corollary 1 thus provides a further adjustment that is motivated by robustness concerns. It is easy to generalize this formula to account for conditioning covariates whose distribution is unspecified, as in Lemma 2.

It is interesting to compute the limit of the MSE-minimizing h function as ϵ tends to infinity in the case where H_π is invertible. This leads to the following expression, which is identical to (13),

$$\lim_{\epsilon \to \infty} h_\epsilon^{\mathrm{MMSE}}(y) = s_{\beta\gamma}(y)'H_{\beta\gamma}^{-1}\nabla_{\beta\gamma}\delta + \widetilde{s}_\pi(y)'\widetilde{\mathbb{H}}_\pi^\dagger\widetilde{\nabla}_\pi\delta, \tag{22}$$

where $\widetilde{\mathbb{H}}_\pi^\dagger$ denotes the Moore–Penrose generalized inverse of $\widetilde{\mathbb{H}}_\pi$. Comparing (22) and Corollary 1 shows that the optimal $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is a Ridge-regularized version of the one-step full MLE, where $(\epsilon n)^{-1}I$ regularizes the projected Hessian matrix $\widetilde{H}_\pi = \widetilde{\mathbb{H}}_\pi\Omega^{-1}$. Our "robust" adjustment remains well-defined under singularity, and it accounts for small or zero eigenvalues of the Hessian in an MSE-optimal way.

A LINEAR REGRESSION EXAMPLE. Studying a linear regression model helps to illustrate some of the main features of our approach. Consider the model

$$Y = X'\beta + U, \qquad X = CZ + V,$$

where Y is a scalar outcome, and X and Z are random vectors of covariates and instruments, respectively, β is a $\dim X$ parameter vector, and C is a $\dim X \times \dim Z$ matrix. We

assume that $U = \pi'V + \xi$, where ξ is normal with zero mean and variance σ^2, independent of X, Z, V, and V is normal with zero mean and nonsingular covariance matrix Σ_V, independent of Z. Let Σ_Z be the covariance matrix of Z, and let $\Sigma_X = C\Sigma_Z C' + \Sigma_V$. For simplicity, we assume that C, Σ_V, Σ_Z, and σ^2 are known, and we take $\Omega = I$ to be the identity matrix. The parameters are thus β and π. In the reference model, we take $\pi = 0$, hence treating X as exogenous whereas the larger model allows for endogeneity. The target parameter is $\delta_{\beta_0, \pi_0} = c'\beta_0$, for a known dim $\beta \times 1$ vector c.

From (21), we have[12]

$$
\begin{aligned}
h_\epsilon^{\text{MMSE}}(y, x, z) = {} & (y - x'\beta_0)x'\Sigma_X^{-1}c \\
& - (y - x'\beta_0)\big[(x - Cz) - \Sigma_V\Sigma_X^{-1}x\big]' \\
& \times \big[(\Sigma_V - \Sigma_V\Sigma_X^{-1}\Sigma_V) + (\epsilon n)^{-1}\sigma^2 I\big]^{-1}\Sigma_V\Sigma_X^{-1}c.
\end{aligned}
\tag{23}
$$

Hence, when $\epsilon = 0$ the minimum-MSE estimator of $c'\beta_0$ is the "one-step efficient" adjustment in the direction of the OLS estimator, with influence function $h_0^{\text{MMSE}}(y, x, z) = (y - x'\beta_0)x'\Sigma_X^{-1}c$. As ϵ tends to infinity, assuming $C\Sigma_Z C'$ is invertible, it follows from (23) that

$$
\lim_{\epsilon \to \infty} h_\epsilon^{\text{MMSE}}(y, x, z) = (y - x'\beta_0)[Cz]'[C\Sigma_Z C']^{-1}c,
$$

which is the influence function of the IV estimator.

For given $\epsilon > 0$ and n, our adjustment remains well-defined even when $C\Sigma_Z C'$ is singular.[13] When $c'\beta_0$ is identified (i.e., when c belongs to the range of C), the minimum-MSE estimator remains well behaved as ϵn tends to infinity, otherwise setting a finite ϵ value is needed to control the increase in variance. The term $(\epsilon n)^{-1}$ in (23) acts as a form of regularization, akin to Ridge regression. In Appendix S4, we show how to extend the parametric setting of this subsection to models defined by moment restrictions, and we revisit this example while dropping the normality assumptions.

3.4 *Semiparametric mixture models*

We now consider a class of semiparametric models, where the distribution of outcomes Y conditional on unobserved latent variables $A \in \mathcal{A}$ is described parametrically by $Y|A \sim g_{\beta_0}(\cdot|A)$, with finite-dimensional unknown parameter $\beta_0 \in \mathcal{B}$, while the distribution of $A \sim \pi_0$ is left unrestricted in the "large" correctly specified model. Here, Π is the set of probability distributions over $\mathcal{A}$. The distribution of observed outcomes as a function of the unknown parameters $\beta_0 \in \mathcal{B}$ and $\pi_0 \in \Pi$ is given by

$$
f_{\beta_0, \pi_0}(y) = \int_{\mathcal{A}} g_{\beta_0}(y|a)\pi_0(a)\, da.
\tag{24}
$$

---

[12]In this case, there is no γ parameter, $s_\beta(y, x|z) = \frac{1}{\sigma^2}x(y - x'\beta_0)$, $s_\pi(y, x|z) = \frac{1}{\sigma^2}(x - Cz)(y - x'\beta_0)$, $\mathbb{E}_{f_Z}H_\beta(Z) = \frac{1}{\sigma^2}\Sigma_X$, $\widetilde{\nabla}_\pi = \nabla_\pi - \Sigma_V\Sigma_X^{-1}\nabla_\beta$, and $\mathbb{E}_{f_Z}\widetilde{H}_\pi(Z) = \frac{1}{\sigma^2}(\Sigma_V - \Sigma_V\Sigma_X^{-1}\Sigma_V)$.

[13]Note that in the absence of an instrument Z, the minimum-MSE estimator coincides with the one-step efficient adjustment in the direction of the OLS estimator.

The parameter of interest is a functional of β_0 and π_0, which takes the form of an expectation over A; that is,

$$\delta_{\beta_0, \pi_0} = \mathbb{E}_{\pi_0} \Delta_{\beta_0}(A) = \int_{\mathcal{A}} \Delta_{\beta_0}(a) \pi_0(a) \, da,$$

where $\Delta_{\beta_0}(a)$ is a known function of β_0 and a.

In Section 6, we will illustrate this setup in two binary choice models: a cross-sectional model, and a dynamic panel data model. In the first case, A is an error term independent of covariates, normally distributed under the reference model. In the second case, A is a latent individual effect correlated with initial conditions, specified using a parametric correlated random-effects reference model (Chamberlain (1984)). In both models, we will estimate average effects, which are expectations with respect to the distribution of A. Our approach will provide insurance against misspecification of the parametric functional forms.

Let us specify a parametric reference model for the distribution of the latent variables A, and denote the reference density by $\pi(\gamma)$, where γ is a finite-dimensional parameter. Under the reference model, the distribution of outcomes is given by $f_{\beta_0, \pi(\gamma_*)}(y) = \int_{\mathcal{A}} g_{\beta_0}(y|a) \pi(a|\gamma_*) \, da$. However, this model may be misspecified, and we assume that the true distribution π_0 belongs to the neighborhood $\Gamma_\epsilon(\gamma_*) = \{\pi_0 \in \Pi : d(\pi_0, \pi(\gamma_*)) \leq \epsilon\}$, which we define here in terms of the Kullback–Leibler (KL) divergence $d(\pi_0, \pi(\gamma_*)) = 2\mathbb{E}_{\pi_0} \log[\pi_0(A)/\pi(A|\gamma_*)]$.

We are going to derive the expression of the minimum-MSE estimator by applying (19). In this setting, elements of the cotangent space $\mathcal{T}$ of $\pi(\gamma)$ at γ_* can be represented by functions $u : \mathcal{A} \mapsto \mathbb{R}$. For example, the gradient $\nabla_\pi \delta_{\beta, \pi}$ is a cotangent element, which can be represented by the function $\Delta_\beta(\cdot)$.[14] For elements $u, w \in \mathcal{T}$, we define their inner product by $u^\top w = \mathrm{Cov}_{\pi(\gamma_*)}[u(A), w(A)]$; that is, the corresponding squared norm in (6) is $\|u\|_{\gamma_*}^2 = \mathrm{Var}_{\pi(\gamma_*)}[u(A)]$. One can show that this norm is indeed the dual to a suitable local approximation to the KL divergence as defined in (6); see Appendix S2.

Let us omit again parameter subscripts from the notation for conciseness. From (24), we see that $s_\pi(y) = \nabla_\pi \log f(y)$ can be represented by the function $g(y|a)/f(y)$. As a result, for $u \in \mathcal{T}$ we have

$$s_\pi(y)^\top u = \mathrm{Cov}\left[u(A), \frac{g(y|A)}{f(y)}\right] = \mathbb{E}[u(A)|Y = y] - \mathbb{E}u(A),$$

where we have used that $\mathbb{E}g(y|A)/f(y) = 1$. In addition, we have

$$\tilde{s}_\pi(y)^\top u = \mathbb{E}[u(A)|Y = y] - \mathbb{E}u(A) - s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \mathbb{E}[s_{\beta\gamma}(Y)u(A)],$$

and, for any function h, $\mathbb{E}[h(Y)s_\pi(Y)]$ can be represented by the function $\mathbb{E}[h(Y)|A = a]$.

Rewriting the first-order condition in equation (19), we thus obtain the following result, which shows that $h_\epsilon^{\mathrm{MMSE}}$ is the solution to a linear system.

---

[14]Note that, since π integrates to one (and therefore tangent space elements integrate to zero), one can equivalently represent $\nabla_\pi \delta_{\beta, \pi}$ as $\Delta_\beta(\cdot) - c$ for any constant c. A possible choice is $c = \mathbb{E}_{\pi(\gamma_*)} \Delta_\beta(A)$.

COROLLARY 2 (Semiparametric mixture models).

$$h_\epsilon^{\text{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma} \delta + (\epsilon n)\{\mathbb{E}[\Delta(A) - \delta - \overline{h}_\epsilon^{\text{MMSE}}(A)|Y = y]$$
$$- s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \mathbb{E}[s_{\beta\gamma}(Y)(\Delta(A) - \overline{h}_\epsilon^{\text{MMSE}}(A))]\},$$

where $\overline{h}_\epsilon^{\text{MMSE}}(a) := \mathbb{E}[h_\epsilon^{\text{MMSE}}(Y)|A = a]$.

Corollary 2 can be generalized to account for conditioning covariates X. We now apply Lemma 2 to provide two generalizations, which we will use in the two examples in Section 6. In the first one, we assume that A and X are independent under π_0. This is the case in our cross-sectional illustration in Section 6.1, where A is an error term independent of X. In this case, π_0 is the marginal distribution of A. We then have the following characterization.

COROLLARY 3 (Semiparametric mixture models, independent covariates).

$$h_\epsilon^{\text{MMSE}}(y, x)$$
$$= \delta(x) - \mathbb{E}_{f_X}\delta(X) + s_{\beta\gamma}(y|x)'\big[\mathbb{E}_{f_X} H_{\beta\gamma}(X)\big]^{-1}\mathbb{E}_{f_X}\nabla_{\beta\gamma}\delta(X)$$
$$+ (\epsilon n)\{\mathbb{E}[\mathbb{E}_{f_X}[\Delta(A, X)] - \mathbb{E}_{f_X}\delta(X) - \overline{h}_\epsilon^{\text{MMSE}}(A)|Y = y, X = x]$$
$$- s_{\beta\gamma}(y|x)'\big[\mathbb{E}_{f_X} H_{\beta\gamma}(X)\big]^{-1}\mathbb{E}_{f_X}\mathbb{E}[s_{\beta\gamma}(Y|X)(\mathbb{E}_{f_X}[\Delta(A, X)] - \overline{h}_\epsilon^{\text{MMSE}}(A))]\},$$

where here $\overline{h}_\epsilon^{\text{MMSE}}(a) := \mathbb{E}_{f_X}[\mathbb{E}(h_\epsilon^{\text{MMSE}}(Y, X)|A = a, X)]$.

In the second generalization, we leave the joint distribution of (A, X) unrestricted under π_0. This is the case in our panel data illustration in Section 6.2, where A is an individual effect that may be correlated with X. In this case, π_0 is the conditional distribution of A given X, and we measure the distance between conditional distributions using $d(\pi_0, \pi(\gamma_*)) = 2\mathbb{E}_{f_X}\mathbb{E}_{\pi_0}\log[\pi_0(A|X)/\pi(A|X, \gamma_*)]$. We then have the following characterization.

COROLLARY 4 (Semiparametric mixture models, correlated covariates).

$$h_\epsilon^{\text{MMSE}}(y, x)$$
$$= \delta(x) - \mathbb{E}_{f_X}\delta(X) + s_{\beta\gamma}(y|x)'\big[\mathbb{E}_{f_X} H_{\beta\gamma}(X)\big]^{-1}\mathbb{E}_{f_X}\nabla_{\beta\gamma}\delta(X)$$
$$+ (\epsilon n)\{\mathbb{E}[\Delta(A, X) - \mathbb{E}_{f_X}\delta(X) - \overline{h}_\epsilon^{\text{MMSE}}(A, X)|Y = y, X = x]$$
$$- s_{\beta\gamma}(y|x)'\big[\mathbb{E}_{f_X} H_{\beta\gamma}(X)\big]^{-1}\mathbb{E}_{f_X}\mathbb{E}[s_{\beta\gamma}(Y|X)(\Delta(A, X) - \overline{h}_\epsilon^{\text{MMSE}}(A, X))]\},$$

where here $\overline{h}_\epsilon^{\text{MMSE}}(a, x) := \mathbb{E}(h_\epsilon^{\text{MMSE}}(Y, X)|A = a, X = x)$.

To provide intuition about the form of the solution in semiparametric mixture models, let us start by considering a setting where β_0 and γ_* are known to the researcher, while abstracting from covariates for simplicity. Let $\mathbb{E}_{\mathcal{Y}|\mathcal{A}}$ and $\mathbb{E}_{\mathcal{A}|\mathcal{Y}}$ denote the conditional expectation operators of Y given A and A given Y, respectively. Corollary 2 implies that (see Appendix S2 for a derivation)

$$h_\epsilon^{\text{MMSE}} = \mathbb{E}_{\mathcal{A}|\mathcal{Y}}\big[\mathbb{E}_{\mathcal{Y}|\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}|\mathcal{Y}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\big]^{-1}(\Delta - \delta), \tag{25}$$

where $\circ$ denotes the composition operator, and $\mathbb{I}_{\mathcal{A}}$ denotes the identity operator; that is, $\mathbb{I}_{\mathcal{A}}\pi = \pi$ for $\pi : \mathcal{A} \to \mathbb{R}$. In semiparametric mixture settings such as panel data models, average effects are often only partially identified or not root-n estimable due to ill-posedness.[15] The presence of the Tikhonov penalty $(\epsilon n)^{-1}$ in (25) bypasses these issues by making the operator $[\mathbb{E}_{\mathcal{Y}|\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}|\mathcal{Y}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}]$ nonsingular. By focusing on a shrinking neighborhood of the reference distribution, as opposed to entertaining any possible distribution, our approach avoids issues of nonidentification and ill-posedness while guaranteeing MSE-optimality within that neighborhood.[16]

Next, consider the estimation of $c'\beta_0$, for c a dim $\beta \times 1$ vector, and assume γ_* known for simplicity. Let $\mathbb{I}_{\mathcal{Y}}h = h$ for $h : \mathcal{Y} \to \mathbb{R}$, and let

$$\mathbb{W}^\epsilon = \mathbb{I}_{\mathcal{Y}} - \mathbb{E}_{\mathcal{A}|\mathcal{Y}}\big[\mathbb{E}_{\mathcal{Y}|\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}|\mathcal{Y}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\big]^{-1}\mathbb{E}_{\mathcal{Y}|\mathcal{A}}.$$

It follows from Corollary 2 that

$$h_\epsilon^{\text{MMSE}}(y) = \mathbb{W}^\epsilon s_\beta(y)'\big\{\mathbb{E}\big[s_\beta(Y)\mathbb{W}^\epsilon s_\beta(Y)'\big]\big\}^{-1}c. \tag{26}$$

As ϵ tends to infinity, $\mathbb{W}^\epsilon$ approximates the functional differencing projection operator $\mathbb{W} = \mathbb{I}_{\mathcal{Y}} - \mathbb{E}_{\mathcal{A}|\mathcal{Y}}\mathbb{E}_{\mathcal{A}|\mathcal{Y}}^\dagger$, where $\mathbb{E}_{\mathcal{A}|\mathcal{Y}}^\dagger$ denotes the Moore–Penrose generalized inverse of $\mathbb{E}_{\mathcal{A}|\mathcal{Y}}$ (see Bonhomme (2012)). In this limit, the minimum-MSE estimator is the one-step approximation to the semiparametric efficient estimator of $c'\beta_0$. Yet, the efficient estimator fails to exist when the matrix denominator in (26) is singular.[17] Here, the term $(\epsilon n)^{-1}$ acts as a regularization of the functional differencing projection, which makes h_ϵ^{MMSE} well-defined irrespective of the nature of identification.

Lastly, consider a model with covariates X that are independent of the latent variables A, as in our illustration in Section 6.1. Assuming that β_0 and γ_* are known, and that $\Delta(A)$ does not depend on X, and letting $\mathbb{E}_{\mathcal{Y},\mathcal{X}|\mathcal{A}}$ and $\mathbb{E}_{\mathcal{A}|\mathcal{Y},\mathcal{X}}$ denote the conditional expectation operators of (Y, X) given A and A given (Y, X), respectively, Corollary 3 implies

$$h_\epsilon^{\text{MMSE}} = \mathbb{E}_{\mathcal{A}|\mathcal{Y},\mathcal{X}}\big[\mathbb{E}_{\mathcal{Y},\mathcal{X}|\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}|\mathcal{Y},\mathcal{X}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\big]^{-1}\mathbb{E}_{f_X}(\Delta - \delta). \tag{27}$$

---

[15]See, for example, Chernozhukov et al. (2013), Pakes and Porter (2013), Severini and Tripathi (2012), and Bonhomme and Davezies (2017).

[16]Note that regular estimation is possible when there exists a function $\psi(y)$ such that $\Delta(A) - \delta = \mathbb{E}[\psi(Y)|A]$. In this case, $\lim_{\epsilon \to \infty} \widehat{\delta}_\epsilon^{\text{MMSE}}$ is consistent for $\mathbb{E}_{\pi_0}\Delta(A) = \delta + \mathbb{E}_{\pi_0}\mathbb{E}[\psi(Y)|A]$ for all π_0.

[17]In discrete choice panel data models, common parameters are generally not point-identified (Chamberlain (2010), Honoré and Tamer (2006)). In panel data models with continuous outcomes, identification and regularity require high-level "nonsurjectivity" conditions, which may be hard to verify (Bonhomme (2012)).

The solution is similar to (25), with the difference that here, due to independence, both Y and X are informative about the latent A.

3.5 *Implementation in parametric models and semiparametric mixture models*

To implement the method in parametric settings, the researcher needs to compute the score and Hessian of the larger model. Since we focus on smooth models, methods based on numerical derivatives or simulation-based approximations can be used. Minimum-MSE estimators are generally not available in closed form in semiparametric mixture models. Nevertheless, in these models $h_\epsilon^{\mathrm{MMSE}}$ can be computed by minimizing the quadratic objective (abstracting from covariates for simplicity)

$$\mathbb{E}\big(\Delta(A) - \delta - \mathbb{E}[h(Y)|A]\big)^2 + (\epsilon n)^{-1}\mathbb{E}h^2(Y), \tag{28}$$

with respect to h, subject to the linear constraints $\mathbb{E}h(Y) = 0$ and $\mathbb{E}h(Y)s_{\beta\gamma}(Y) = \nabla_{\beta\gamma}\delta$. This is a regularized linear inverse problem (see, e.g., Engl et al. (2000), and Kress (2014)), which is well-posed given the presence of the Tikhonov penalty $(\epsilon n)^{-1}$. Numerous numerical approaches have been developed to solve linear inverse problems. In the illustrations, we implement a simulation-based method that relies on matrix operations. In Appendix S3, we describe this method, and also explain how we compute confidence intervals. Note that, given initial estimates $\widehat{\beta}$ and $\widehat{\gamma}$, computing minimum-MSE estimators and confidence intervals does not require nonlinear optimization.

4. USING MINIMUM-MSE ESTIMATES AND CONFIDENCE INTERVALS FOR SENSITIVITY ANALYSIS

In this section, we discuss how to apply our approach in practice. So far, we have shown how to compute minimum-MSE estimators and confidence intervals for different values of ϵ. We now describe a strategy to choose an interpretable range for ϵ, and report the results of the estimation on this range, in the spirit of sensitivity analysis.

4.1 *Context-specific interpretations of magnitudes of ϵ*

In order to apply our approach, an important first step is to form intuition about orders of magnitudes of ϵ in the model under study. The literature on sensitivity analysis that we referred to in the introduction provides intuition in certain models; see, for example, the analysis of linear IV models in Conley et al. (2012). In Sections 5 and 6, we will discuss the magnitudes of ϵ in our examples. In our evaluation of a conditional cash transfer program in Section 5, misspecification stems from omitted stigma effects in households' preferences. In this case, we will show that ϵ can be mapped to the ratio of the marginal utility of the subsidy (i.e., the "stigma" effect) to that of consumption. Economic intuition can then suggest whether a given ϵ is large or small. In this setting, one can motivate the assumption that ϵ shrinks as n increases as reflecting that the econometrician's uncertainty about the presence of stigma effects diminishes when the sample gets larger.

In the binary choice models we study in Section 6, π is infinite-dimensional, and $\epsilon/2$ is the squared radius of a Kullback–Leibler ball around a normal density. To visualize the implications of assuming that the true π belongs to an ϵ-neighborhood of the normal, we will plot worst-case probability bounds, and show how setting ϵ to a particular value imposes ex ante restrictions on the parameter of interest. Local approximations to bounds on functionals of π are easy to compute. Alternatively, one may report estimated worst-case distributions—that is, a π_0 that achieves the supremum in (7)—in the spirit of Christensen and Connault (2019).

4.2 *Interpretation based on statistical testing*

As a complement to forming context-specific intuition about magnitudes, here we outline a generic interpretation based on statistical testing. Specifically, we show that setting ϵ is isomorphic to setting a lower bound on the local power of a likelihood-ratio test of the reference model, against alternatives outside the neighborhood $\Gamma_\epsilon(\gamma_*)$ in certain directions. The ϵ-neighborhood will thus contain all models that are hard to statistically distinguish from the reference model in those directions. This logic has antecedents in robust statistics (Huber and Ronchetti (2009)), and robust control in economics (Hansen and Sargent (2008)).

To proceed, let us focus on the parametric case of Section 3.3 with identity weight matrix Ω. Let v be a unitary vector, and consider a likelihood-ratio test of the null hypothesis $H_0 : \pi_0 = \pi(\gamma_*)$ against the local alternative $H_1 : \pi_0 = \pi(\gamma_*) + \xi v/\sqrt{n}$, for some constant $\xi > 0$. Let the size of the test be $\alpha \in (0, 1)$. The local power of the test is then $p = \Pr(Z[\mu] > \widetilde{c}_\alpha)$, where $\widetilde{c}_\alpha$ is the $(1 - \alpha)$-quantile of the chi-squared distribution with one degree of freedom, and $Z[\mu]$ follows a noncentral chi-squared distribution with one degree of freedom and noncentrality parameter $\mu = \|\widetilde{H}_\pi^{1/2} v\| \xi$; see, for example, Van der Vaart (2007, p. 237).[18] For given α and p values, let $\mu(\alpha, p)$ be such that $\Pr(Z[\mu(\alpha, p)] > \widetilde{c}_\alpha) = p$.[19] It follows that $\mu(\alpha, p) = \|\widetilde{H}_\pi^{1/2} v\| \xi$. Hence, noting that $\mu(\alpha, p)$ is increasing in p, and defining

$$\epsilon(v) = \frac{\mu(\alpha, p)^2}{nv'\widetilde{H}_\pi v},$$

taking $\epsilon \geq \epsilon(v)$ ensures that local power in direction v is at least p whenever $\xi/\sqrt{n} \geq \epsilon^{\frac{1}{2}}$.[20] Note that, for fixed α and p, the product $\epsilon(v)n$ is independent of n.

To ensure power larger than p outside the neighborhood in *all* directions v, one could compute the *supremum* of $\epsilon(v)$ over all directions.[21] Setting ϵ larger than all $\epsilon(v)$'s

---

[18]Here, $\widetilde{H}_\pi$ is the usual parametric (projected) Hessian matrix, since Ω is the identity.

[19]$\mu(\alpha, p)$ is implicitly defined by $\Phi(\mu(\alpha, p) + \Phi^{-1}(\alpha/2)) + \Phi(-\mu(\alpha, p) + \Phi^{-1}(\alpha/2)) = p$, where Φ is the standard normal cumulative distribution function.

[20]This definition is easy to extend to the general locally quadratic case of Section 3.1. Let v be a unitary direction in the tangent space $\overline{\mathcal{T}}$ of $\pi(\gamma)$ at γ_*, and let $\Omega_{\gamma_*} : \overline{\mathcal{T}} \to \mathcal{T}$ be the linear operator defined in the Appendix in the Online Supplementary Material. In the parametric case Ω_{γ_*} is simply the matrix Ω. In the general setup the noncentrality parameter is $\langle v, \widetilde{H}_\pi \Omega_{\gamma_*} v\rangle^{1/2}\xi$, and $\epsilon(v) = \mu(\alpha, p)^2/(n\langle v, \widetilde{H}_\pi \Omega_{\gamma_*} v\rangle)$, for $\langle v, u\rangle \in \mathbb{R}$ the scalar product between $v \in \overline{\mathcal{T}}$ and $u \in \mathcal{T}$.

[21]It is sufficient to consider directions that are orthogonal to the directions $\nabla_{\gamma_*}\pi'$ of the reference model.

is motivated by a desire to calibrate the fear of misspecification of the researcher: when p is large, say 80%, all alternatives outside the neighborhood $\Gamma_\epsilon(\gamma_*)$ are then easy to statistically distinguish from the reference model based on a sample of n observations. However, for the supremum of $\epsilon(v)$ to be finite, $\widetilde{H}_\pi$ needs to be nonsingular, which precludes models with partial or irregular identification. As an example, in the linear model of Section 3.3, the supremum of $\epsilon(v)$ is infinite whenever $\Sigma_X - \Sigma_V = C\Sigma_Z C'$ is singular; that is, whenever the IV model is underidentified. In such a case, there thus exist certain directions along which the specification test has no power, no matter how large ϵ is. Likewise, in semiparametric models, the eigenvalues of the infinite-dimensional operator $\widetilde{H}_\pi$ may not be bounded away from zero due to ill-posedness.

In partially or irregularly identified models, given some fixed values of α and p, a possibility is to report several ϵ value: a *first* value ϵ_1 that corresponds to the *infimum* of $\epsilon(v)$ over all directions v—hence to the most favorable direction; a *second* value $\epsilon_2 \geq \epsilon_1$, such that power is at least p outside the neighborhood in the most favorable direction in the subspace of directions orthogonal to the most favorable one; a *third* value $\epsilon_3 \geq \epsilon_2$ that provides power guarantees along the most favorable direction orthogonal to the previous two ones, and so on. Letting $\lambda_k(B)$ denote the kth largest eigenvalue of B, we have

$$\epsilon_k = \frac{\mu(\alpha, p)^2}{n\lambda_k(\widetilde{H}_\pi)}, \quad \text{for } k = 1, 2, \ldots \tag{29}$$

We will report the first few ϵ_k values in our illustrations—taking $\alpha = 5\%$ and $p = 80\%$—as a complement to context-specific interpretations of orders of magnitude.[22]

4.3 *Reporting results*

By providing intuition about ϵ, either through an interpretation of magnitudes in the context under study (see Section 4.1), and/or through a generic approach based on statistical testing (see Section 4.2), the researcher selects a range of possible values for ϵ. We then recommend plotting the minimum-MSE estimator, and its associated 95% confidence interval, as a function of ϵ on this range. In our illustrations, we will use this device to report results, and we will indicate particular values of ϵ on the x-axis to facilitate interpretation.

By reporting those minimum-MSE estimates and bias-adjusted confidence intervals, we learn about the fragility of the estimation results under the reference model parameterized by γ, relative to the larger model parameterized by π. Exploring the sensitivity to model misspecification in that way is in line with the traditional suggestion of comparing estimation results obtained from different model specifications (see, e.g., Leamer (1983, 1985)). However, our local approach does not require the researcher to estimate the larger model, which is particularly relevant in situations where the latter is partially or irregularly identified, or computationally hard to estimate.

---

[22]In Appendix S3, we describe how to compute ϵ_k in semiparametric mixture models using a simulation-based approach.

4.4 *Remark: Shape of neighborhoods and choice of norm*

Implementing our approach requires choosing a norm on Π, which governs the shape of $\Gamma_\epsilon(\gamma_*)$. In parametric models, the researcher may have a preferred weight matrix Ω, thus putting more weight on certain elements of the vector π. An automatic weighting scheme is to set Ω to be equal to the diagonal of the projected Hessian matrix $\widetilde{\mathbb{H}}_\pi$. This choice can be motivated using a statistical testing logic as in Section 4.2, focusing on componentwise directions in the canonical basis of $\mathbb{R}^{\dim \pi}$. Taking the diagonal, instead of the entire matrix $\widetilde{\mathbb{H}}_\pi$, as a weight is in line with our aim to cover models where the parameter of interest may not be regularly estimable.[23]

In semiparametric mixture models, where Π is a set of densities, we rely on the Kullback–Leibler divergence for computational convenience. KL is locally quadratic, and this choice allows us to obtain the explicit characterizations of Lemma 1, and to compute minimum-MSE estimators by solving linear systems. Note, however, that the KL divergence does not impose shape or smoothness restrictions on the densities inside the neighborhood.

5. Empirical illustration: Conditional cash transfers in Mexico

The goal of this section is to predict program impacts in the context of the PROGRESA conditional cash transfer program, building on the structural evaluation of the program in Todd and Wolpin (2006, TW hereafter) and Attanasio et al. (2012, AMS). We estimate a simple model in the spirit of TW, and adjust its predictions against a specific form of misspecification under which the program may have a "stigma" effect on preferences.

5.1 *Setup*

Following TW and AMS, we focus on PROGRESA's education component, which consists of cash transfers to families conditional on children attending school. Those represent substantial amounts as a share of total household income. The implementation of the policy was preceded by a village-level randomized evaluation in 1997–1998. As TW and AMS point out, the randomized control trial is silent about the effect that other, related policies could have, such as higher subsidies or unconditional income transfers, which motivates the use of structural methods.

To analyze this question, we consider a simplified version of TW's model described in Wolpin (2013), which is a static, one-child model with no fertility decision. To describe this model, let $U(C, S, \tau, v)$ denote the utility of a unitary household, where C is consumption, $S \in \{0, 1\}$ denotes the schooling attendance of the child, τ is the level of the PROGRESA subsidy, and v are taste shocks. Utility may also depend on characteristics X, which we abstract from for conciseness in the presentation.[24] Note the direct presence

---

[23]In applications, other norms may have particular appeal. For example, measuring deviations according to the supremum norm will lead to an ℓ^1 dual norm in (10), in the spirit of Armstrong and Kolesár (2021). While our estimators and confidence intervals remain well-defined in this case, that setting is not locally quadratic.

[24]Empirically, we include as covariates the age of the child and her parents, distance to the nearest school, eligibility and year indicators, and the highest grade obtained. We perform estimation separately by gender.

of the subsidy τ in the utility function, which may reflect a stigma effect. This direct effect plays a key role in the analysis. The budget constraint is: $C = Y + W(1 - S) + \tau S$, where Y is household income and W is the child's wage. This is equivalent to: $C = Y + \tau + (W - \tau)(1 - S)$. Hence, in the absence of a direct effect on utility, the program's impact is equivalent to an increase in income and a decrease in the child's wage.

Following Wolpin (2013), we parameterize the utility function as

$$U(C, S, \tau, v) = aC + bS + dCS + \lambda\tau S + Sv,$$

where λ denotes the direct (stigma) effect of the program. The schooling decision is then

$$S = \mathbf{1}\{U(Y + \tau, 1, \tau, v) > U(Y + W, 0, 0, v)\}$$
$$= \mathbf{1}\{v > a(Y + W) - (a + d)(Y + \tau) - \lambda\tau - b\}.$$

Assuming that v is standard normal, independent of wages, income, and program status (i.e., of the subsidy τ), we obtain

$$\Pr(S = 1 | y, w, \tau) = 1 - \Phi[a(y + w) - (a + d)(y + \tau) - \lambda\tau - b],$$

where Φ is the standard normal cdf.

We use the specification with $\lambda = 0$ as our reference model, and estimate it on control villages. When $\lambda = 0$, the average effect of the subsidy on school attendance is

$$\mathbb{E}\big[\Pr(S = 1 | Y, W, \tau = \tau^{\text{treat}}) - \Pr(S = 1 | Y, W, \tau = 0)\big]$$
$$= \mathbb{E}\big(\Phi[a(Y + W) - (a + d)(Y + \tau^{\text{treat}}) - b] - \Phi[a(Y + W) - (a + d)Y - b]\big).$$

As Wolpin (2013) notes, data under the subsidy regime ($\tau = \tau^{\text{treat}}$) is not needed to construct an empirical counterpart to this quantity, since treatment status is independent of Y, W.[25]

We contrast two strategies to predict the effect of the program and other counterfactual policies, while accounting for misspecification of the reference model due to the presence of stigma effects. The first strategy—which we refer to as ex ante policy prediction—is only based on data from control villages, whereas the second strategy—ex post prediction—combines both control and treated villages. In both cases, we allow for $\lambda \neq 0$ in the larger model. While in the present simple static context one could easily estimate a version of the larger model, in dynamic structural models such as the one estimated by TW, estimating a different model in order to assess the impact of any given form of misspecification may be computationally prohibitive. This highlights an advantage of our approach, which does not require the researcher to estimate the parameters under a new model.

To cast this setting into our framework, let $\beta = (a, b, d)$, $\pi = \lambda$, and

$$\delta_{\beta, \pi} = \mathbb{E}\big(\Phi[a(Y + W) - (a + d)(Y + \tau^{\text{treat}}) - \lambda\tau^{\text{treat}} - b] - \Phi[a(Y + W) - (a + d)Y - b]\big).$$

---

[25]AMS make a related point (albeit in a different model), and use both control and treated villages to estimate their structural model. AMS also document the presence of general equilibrium effects of the program on wages. We abstract from such effects in our analysis.

We focus on the effect on eligible (i.e., poorer) households. We add covariates to gender-specific school attendance equations, which include the age of the child and her parents, year indicators, distance to school, and an eligibility indicator. We report estimates of δ_{β_0, π_0} as well as confidence intervals.

5.2 *Empirical results*

We use the sample from TW. We drop observations with missing household income, and focus on 1219 boys and 1089 girls aged 12 to 15.[26] Descriptive statistics on the sample show that average weekly household income is 242 pesos, the average weekly wage is 132 pesos, and the PROGRESA subsidy ranges between 31 and 59 pesos per week depending on grade and gender. Average school attendance drops from 90% at age 12 to between 40% and 50% at age 15.

We start by providing intuition regarding the values of ϵ in the present context. In the structural model, λ is the marginal utility of the subsidy for households sending their child to school. In turn, the marginal utility of consumption is given by $a + d$. Hence, bounding λ^2 by ϵ is equivalent to bounding the ratio of marginal utility of the subsidy to marginal utility of consumption by $\sqrt{\epsilon}/(a + d)$. For example, households valuing the subsidy as much as consumption in absolute value—arguably an upper bound on the stigma effect—corresponds to $\overline{\epsilon} = (a + d)^2$.

In Figure 1, we show the minimum-MSE estimator of the impact of the PROGRESA subsidy on eligible households, together with 95% confidence intervals, for a range of values around $\overline{\epsilon} = (a + d)^2$ (which we show in the dashed vertical line). In the horizontal dotted line, we show estimates based on the reference model. In the top panel, we show ex ante prediction results based on control villages only. We see that the minimum-MSE estimator and the one based on the reference model are equal in this case. This is intuitive, since $\pi = \lambda$ is scalar, and control villages provide no information about it.[27] However, the confidence intervals—which account for model misspecification—are large. When $\epsilon = \overline{\epsilon}$, the 95% intervals include zero for both genders. This quantifies the uncertainty associated with ex ante prediction when the researcher does not rule out the presence of stigma.

In the bottom panel of Figure 1, we show the results of ex post prediction based on both control and treated villages. In the sample of boys, the minimum-MSE estimator is lower than the one based on the reference model, suggesting that the reference model is misspecified. In contrast, the two estimators are close to each other in the sample of girls.[28] In addition, the 95% confidence intervals are substantially tighter than when us-

[26]Children's wages are only observed for those who work. We impute potential wages to all children based on a linear regression that in particular exploits province-level variation and variation in distance to the nearest city, similar to AMS.

[27]This is analogous to the case of a linear regression with endogeneity and no instrument, which we mentioned in footnote 13.

[28]Note that, for boys, minimum-MSE estimates at all ϵ values—including $\epsilon = 0$—are lower than the estimate from the reference model (note that here we estimate the reference model using control villages only, and use both controls and treated to compute the minimum-MSE estimator). This suggests that the functional form of the schooling decision is not invariant to treatment status, highlighting that predictions based off control villages are less satisfactory for boys (as also found by TW).
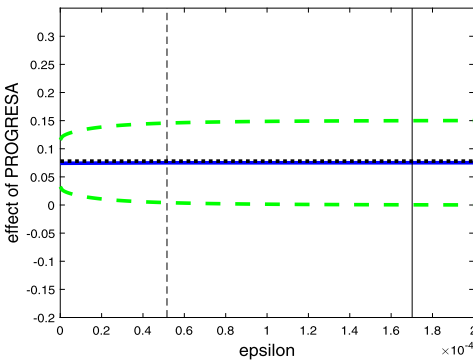
Ex ante

Girls
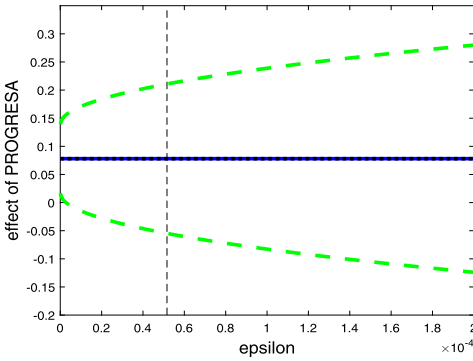
Boys
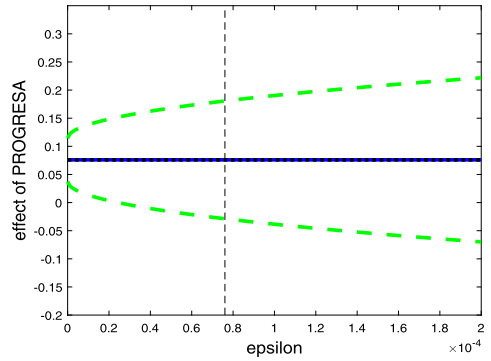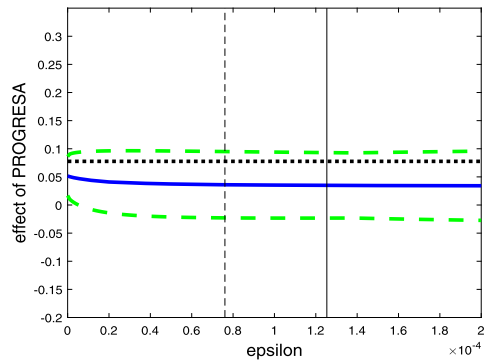
Ex post

Girls

Boys

FIGURE 1. Effect of the PROGRESA subsidy as a function of neighborhood size ϵ. *Notes*: Sample from Todd and Wolpin (2006). In the top panel, we show estimates based on control villages only, in the bottom panel we show estimates based on both controls and treated. We report ϵ on the x-axis. The minimum-MSE estimates of the effect of PROGRESA on school attendance are shown in solid. 95% confidence intervals based on those estimates are in dashed. The dotted line shows the unadjusted prediction based on the reference model. The dashed vertical line indicates $\bar{\epsilon}$ (at which households value consumption and the subsidy equally) and the solid vertical line indicates ϵ_1 (at which a 5% likelihood ratio specification test has power 80%). Girls (left) and boys (right).

ing control villages only. When $\epsilon = \bar{\epsilon}$ (shown in the vertical dashed line), the program estimates on school attendance are positive and marginally significant at the 5% level for girls, and positive and marginally significant at 10% for boys. In the vertical solid line, we highlight the value ϵ_1 given by (29) for $k = 1$. Since $\pi = \lambda$ is scalar, setting $\epsilon \geq \epsilon_1$ ensures that, for all models outside the neighborhood, a 5%-likelihood ratio specification test has local power larger than 80%. Taking $\epsilon = \epsilon_1$ implies that the ratio of marginal

TABLE 1. Effect of the PROGRESA subsidy and counterfactual reforms.

	Ex ante		Ex post	
	PROGRESA impacts			
	Girls	Boys	Girls	Boys
Model-based estimate	0.078	0.076	0.078	0.078
nonrobust CI	(0.017,0.140)	(0.038,0.114)	(0.026,0.130)	(0.043,0.112)
robust CI	(−0.055,0.211)	(−0.029,0.181)	(0.006,0.150)	(0.008,0.148)
MMSE estimate	0.078	0.076	0.075	0.036
robust CI	(−0.055,0.211)	(−0.029,0.181)	(0.004,0.146)	(−0.023,0.095)
Experimental	–	–	0.087	0.050
	Counterfactual 1: doubling subsidy			
	Girls	Boys	Girls	Boys
Model-based estimate	0.142	0.133	0.142	0.136
robust CI	(−0.069,0.353)	(−0.018,0.284)	(0.027,0.258)	(0.033,0.239)
MMSE estimate	0.142	0.133	0.137	0.076
robust CI	(−0.069,0.353)	(−0.018,0.284)	(0.022,0.251)	(−0.017,0.163)
	Counterfactual 2: unconditional transfer			
	Girls	Boys	Girls	Boys
Model-based estimate	0.003	0.005	0.003	0.005
robust CI	(−0.213,0.218)	(−0.231,0.240)	(−0.208,0.214)	(−0.239,0.249)
MMSE estimate	0.003	0.005	0.009	−0.071
robust CI	(−0.213,0.218)	(−0.231,0.240)	(−0.232,0.250)	(−0.279,0.136)

Note: Sample from Todd and Wolpin (2006). In the left two columns we show estimates based on control villages only, in the right two columns we show estimates based on both controls and treated. "Model-based" estimates are based on the reference model. $\epsilon = \bar{\epsilon}$, which corresponds to households valuing consumption and the subsidy equally in absolute value. CI are 95% confidence intervals. The unconditional transfer amounts to 5000 pesos in a year.

utility of the subsidy to marginal utility of consumption is bounded by 1.8 (girls) and 1.3 (boys). While ϵ_1 is larger than $\bar{\epsilon}$, the implied minimum-MSE estimators and confidence intervals are similar.[29]

In Table 1, we report estimates of the program impacts, as well as predictions of counterfactual policies. The left two columns correspond to ex ante prediction based on control villages only, while the right two columns correspond to ex post prediction based on both controls and treated. We show the results for $\epsilon = \bar{\epsilon}$, corresponding to equal marginal utilities of subsidy and consumption in absolute value. In the top panel, we focus on the impact of the PROGRESA subsidy on eligible households. We see that PROGRESA has a positive impact on attendance of both boys and girls. The impacts predicted by the reference model are large, approximately 8 percentage points, and are quite close to the results reported in Todd and Wolpin (2006, 2008). However, the confidence intervals which account for model misspecification (third row, left two columns) are very large for both genders.

When adding treated villages to the sample (right two columns in Table 1), confidence intervals accounting for misspecification are tighter. Moreover, the minimum-MSE point estimates and those based on the reference model differ in this case. For

[29]Note that ϵ_1 is infinite in the ex ante case (top panel of Figure 1). This is due to control villages not providing any information about π in this case.

boys, the minimum-MSE estimate is substantially lower than the one based on the reference model (3.6% versus 7.8%), while for girls the effects are similar. Interestingly, for boys the minimum-MSE estimates are closer to the experimental differences in means between treated and control villages.

Lastly, in the middle and bottom panels of Table 1, we show estimates of the effects of two counterfactual policies: doubling the PROGRESA subsidy, and removing the conditioning of the income transfer on school attendance. Unlike for the main PROGRESA impacts, there is no experimental counterpart to such counterfactuals. While ex ante predictions are associated with wide confidence intervals, ex post minimum-MSE estimates based on both control and treated villages predict a substantial effect of doubling the subsidy on girls' attendance, and a more moderate effect on boys. By contrast, we find insignificant effects of an unconditional income transfer.

6. Numerical illustrations: Binary choice models

In this section, we apply our approach to cross-sectional and panel data binary choice models, where we allow for misspecification of the distribution of unobservables. In both applications there is a substantial amount of misspecification, and we use simulations to assess the behavior of the minimum-MSE estimator—which is theoretically justified under local misspecification—in these settings of practical relevance.

6.1 Cross-sectional binary choice

Consider the binary choice model

$$Y = \mathbb{1}\{X'\beta_0 + A \geq 0\}, \tag{30}$$

where A follows a distribution π_0, independent of X. We are interested in estimating the prediction function $\delta_{\beta_0, \pi_0} = \mathbb{E}_{\pi_0}[\mathbb{1}\{x_0'\beta_0 + A \geq 0\}]$, at some x_0 not necessarily in the support of X. We focus on the reference specification $A \sim \mathcal{N}(0, 1)$, independent of X. We allow for the possibility that this parametric model is misspecified, while maintaining independence between A and X under π_0. We observe an i.i.d. sample (Y_i, X_i) for $i = 1, \ldots, n$.

In neighborhoods that consist of distributions of A independent of X, the minimum-MSE influence function is given by Corollary 3, with $\nabla_\beta \delta = x_0 \phi(x_0'\beta_0)$ for ϕ the standard normal density, $\Delta(a) = \mathbb{1}\{x_0'\beta_0 + a \geq 0\}$, and without γ parameter. Given a preliminary estimator $\widehat{\beta}$ (e.g., obtained by probit), an empirical counterpart to $\overline{h}_\epsilon^{\text{MMSE}}(a)$ is

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i'\widehat{\beta} + a \geq 0\} h_\epsilon^{\text{MMSE}}(1, X_i) + \left(1 - \mathbb{1}\{X_i'\widehat{\beta} + a \geq 0\}\right) h_\epsilon^{\text{MMSE}}(0, X_i).$$

We compute $h_\epsilon^{\text{MMSE}}(1, X_i)$ and $h_\epsilon^{\text{MMSE}}(0, X_i)$, for $i = 1, \ldots, n$, based on Corollary 3 by solving a linear system. In this model, all conditional expectations are available in closed form.
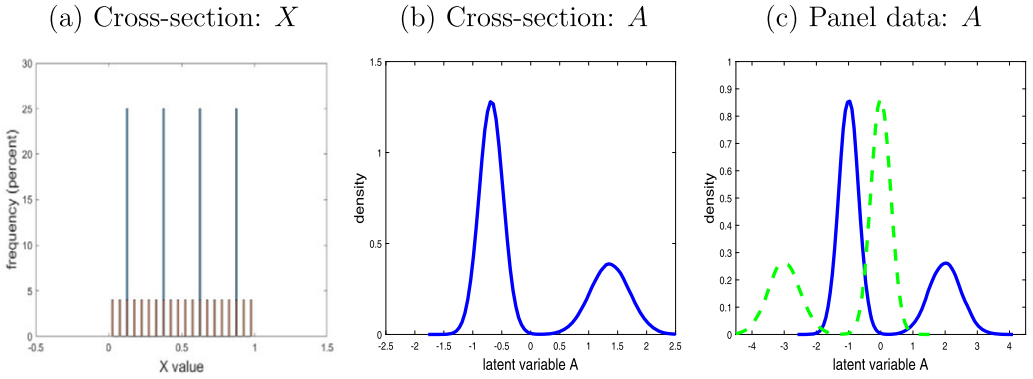
(a) Cross-section: X (b) Cross-section: A (c) Panel data: A

FIGURE 2. Distributions of X and A in the binary choice models. *Notes*: In panel (a), we show the frequencies of covariates (i.e., the first component of X) in the cross-sectional model, and in (b) we show the true density of A for the same model. In (c), we show the true densities of A in the panel data model for $Y_0 = 0$ (in solid) and $Y_0 = 1$ (in dashed).

In model (30), under independence between A and X, β_0 and π_0 are point-identified up to scale under sufficiently rich support of X (Manski (1988)). Under such conditions, δ_{β_0, π_0} is identified. More generally, it is partially identified. We now set up a simulation where the support of X is discrete, and we vary the number of support points and the target x_0. In this way, we learn how our estimators and confidence intervals perform in settings where the support of X, and hence the size of the identified set, vary.

We will show estimates in data generating processes (DGPs) with a scalar covariate and an intercept, and $\beta_0 = (2, -1)'$, where the second element corresponds to the intercept. We draw 1000 simulated samples of size $n = 500$, where A has mean zero and variance one, and is distributed as a mixture of two normals whose means are approximately two standard deviations apart. Covariates are discrete uniform on $[0, 1]$, with either $n_X = 4$ or $n_X = 20$ points of support. We show the densities of X and A in panels (a) and (b) of Figure 2. We focus on the predicted values at $x_0 = (0.5, 1)'$ and $x_0 = (-0.5, 1)'$, respectively. We refer to the first case as *interpolation*, and to the second one as *extrapolation*.

We report minimum-MSE estimates and confidence intervals on a range of ϵ values. To provide intuition about orders of magnitude, one can compute (local approximations to twice) the KL divergence between the standard normal and other common distributions. For example, a scaled student-t distribution with unitary variance and 5, 3, or 2.1 degrees of freedom, respectively, corresponds to a distance of 0.14, 0.24, and 1.5; the true bimodal π_0 in the DGP corresponds to a distance of 1.6; and the standardized logistic corresponds to a distance of 0.07. Restricting π_0 to belong to an ϵ-neighborhood of the normal also has implications for its functionals. As an example, in Figure 3, we show pointwise bounds on $\mathbb{E}_{\pi_0}[\mathbf{1}\{a + A \geq 0\}]$, as a function of $\Phi(a)$, computed using small-ϵ approximations. We see that taking $\epsilon = 0.1$ tightly restricts possible values that the parameter can take. By contrast, when $\epsilon = 1$, the a priori bounds on the parameter are very wide for a close to 0.5—which is relevant for the *interpolation* case—but the
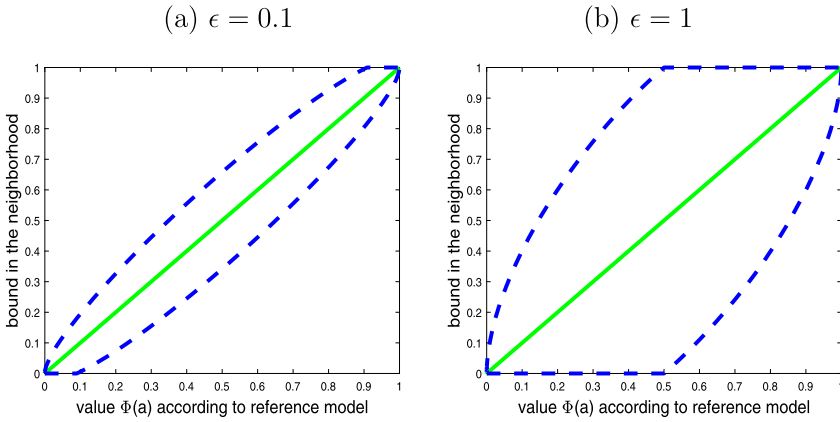
FIGURE 3. Bounds on $\mathbb{E}_{\pi_0}[\mathbf{1}\{a + A \geq 0\}]$ in ϵ-neighborhoods of the standard normal. *Notes*: The dashed lines show $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\pi_0}[\mathbf{1}\{a + A \geq 0\}]$ and $\inf_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\pi_0}[\mathbf{1}\{a + A \geq 0\}]$ as a function of $\Phi(a)$. The 45-degree line is in solid. Computations are based on small-ϵ approximations.

neighborhood does restrict the parameter value when a in close to zero or one—which is relevant for the *extrapolation* case. Given this, we will interpret ϵ values of the order of 0.1 or lower as reflecting "mild" misspecification, and values of the order of 1 or larger as corresponding to "large" misspecification.

In this example, it is also informative to interpret ϵ by relating it to the power of a specification test, as we described in Section 4.2. When X has 4 support points, $\widetilde{H}_\pi$ has only $n_X - 1 = 3$ nonzero eigenvalues (since the $X'\beta$ partition the real line into $n_X + 1$ intervals, and the two elements in β are estimated), two of them corresponding to non-constant eigenfunctions. In this case, ϵ_1 and ϵ_2 given by (29) are approximately 0.2 and 0.4 on average, where we set size to $\alpha = 5\%$ and power to $p = 80\%$. When X has 20 support points, $\widetilde{H}_\pi$ has $n_X - 1 = 19$ nonzero eigenvalues corresponding to nonconstant eigenfunctions. The first three values ϵ_1, ϵ_2, and ϵ_3 are approximately 0.3, 0.6, and 1.1 on average. In contrast with the parametric case of Section 5, here setting $\epsilon \geq \epsilon_k$ only provides power guarantees along particular directions. In Appendix S5, we plot those directions, and we provide additional intuition about the interpretation of ϵ based on statistical testing in this example.

We show the results of the simulation in Figure 4. Consider first the top panel, where we wish to interpolate the prediction function at $x_0 = 0.5$. When X has 4 support points, we see that the probit estimator based on the reference model, indicated by the solid horizontal line, is substantially biased. By contrast, the bias of the minimum-MSE estimator is smaller, and it decreases as ϵ increases. We see that the minimum-MSE estimator is close to unbiased for both ϵ_1 and ϵ_2. Moreover, the dispersion of the estimator is stable as ϵ increases. In addition, we compute the identified set for δ_{β_0, π_0} in the DGP using linear programming and a grid of β_0 values. We find $[0.334, 0.345]$, which shows that the identified set is not wide in this DGP.
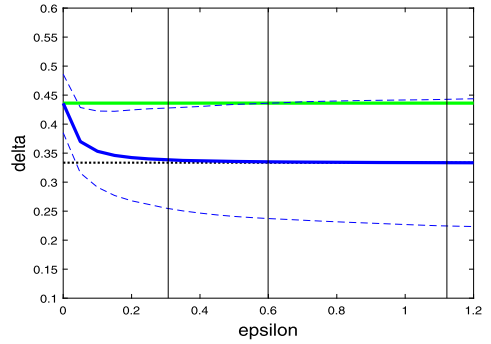
The case where X has 20 support points is overall quite similar, but with several differences. We see that the minimum-MSE estimator is virtually unbiased when $\epsilon \geq \epsilon_1$. In
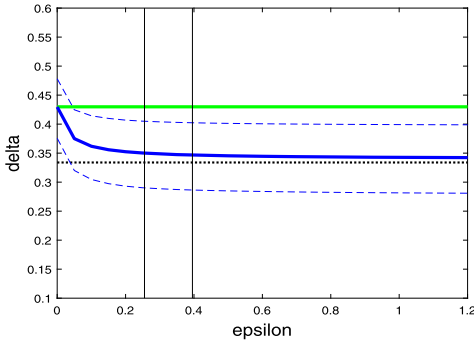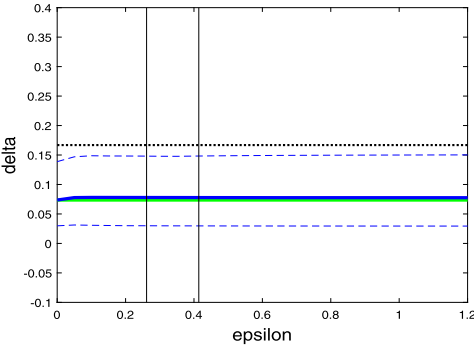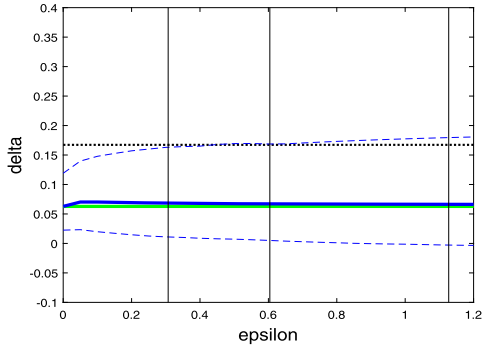
FIGURE 4. Minimum-MSE estimator in the cross-sectional binary choice model. *Notes*: The solid horizontal line corresponds to the mean probit estimator among 1000 simulations, the solid curve to the mean minimum-MSE estimator (with 2.5% and 97.5% percentiles in dashed), and the dotted horizontal line to the truth (i.e., δ_{β_0, π_0}, for the β_0 and π_0 we used in the DGP). ϵ is reported on the x-axis, and the vertical lines indicate ϵ_k given by (29), for $k \in \{1, 2, 3\}$. n_X denotes the number of points of support of the first component of X. $n = 500$.

this case, the identified set for δ_{β_0, π_0} is essentially a singleton: $[0.334, 0.335]$. Moreover, we see that the variance of the minimum-MSE estimator increases with ϵ. Such a variance increase, and the associated regularization role of ϵ, also characterize models with continuously distributed covariates and other ill-posed inverse problems.

Lastly, consider the lower panel in Figure 4. This *extrapolation* case is very different from the *interpolation* one. Indeed, the data provide little information about the value of the prediction function at $x_0 = -0.5$. To illustrate, the identified set for δ_{β_0, π_0} is $[0, 0.3219]$ (resp., $[0, 0.2956]$) when X has 4 (resp., 20) points of support. We see that the minimum-MSE estimator has approximately the same bias as the probit estimator in this case.

We show additional information about the simulation results in Tables S1 and S2 in the Appendix. In particular, we report the lengths of our 95% confidence intervals (CI)

for δ_{β_0, π_0}, which are asymptotically valid under ϵ-misspecification, and the associated coverage probabilities. In all DGPs, we find that, when taking $\epsilon \geq \epsilon_1$, the confidence intervals contain the true value with a probability that exceeds 95%.[30]

6.2 *Dynamic panel data binary choice*

In this subsection, we present simulations in the following dynamic panel data probit model with individual effects

$$Y_t = \mathbb{1}\{\beta_0 Y_{t-1} + A + U_t \geq 0\}, \quad t = 1, \ldots, T, \tag{31}$$

where $U_1, \ldots, U_T$ are i.i.d. standard normal, independent of A and Y_0. Here, Y_0 is observed, so there are effectively $T + 1$ time periods. We focus on the average state dependence effect $\delta_{\beta_0, \pi_0} = \mathbb{E}_{\pi_0}[\Phi(\beta_0 + A) - \Phi(A)]$, and we will also report estimates of the autoregressive parameter β_0. We assume that the probit conditional likelihood given individual effects and lagged outcomes is correctly specified. However, we do not assume knowledge of π_0 or its functional form. We specify a normal reference density for A given Y_0, with mean $\mu_1 + \mu_2 Y_0$ and variance σ^2; hence, here $\gamma = (\mu_1, \mu_2, \sigma^2)'$. Binary choice panel data models are often partially identified for fixed T (Chamberlain (2010), Honoré and Tamer (2006)), and no semiparametrically consistent estimators of β_0 and δ_{β_0, π_0} in the dynamic probit model are available in the literature. Here, we report simulation results suggesting that minimum-MSE estimators can perform well under sizable misspecification of the reference density.

In the simulation, we set a bimodal distribution that has modes $\{-1, 2\}$ when $Y_0 = 0$ and $\{-3, 0\}$ when $Y_0 = 1$, with some asymmetry between the two modes; see panel (c) of Figure 2. We draw Y_0 from a Bernoulli distribution with probability 0.5. We take $n = 500$, and show the results for $T = 5, 10,$ and 20, based on 1000 simulations. In neighborhoods that consist of unrestricted joint distributions π_0 of (A, X), the minimum-MSE h function is given by Corollary 4, for $X = Y_0$, and either $\Delta(a) = \Phi(\beta_0 + a) - \Phi(a)$ or $\Delta(a) = \beta_0$, depending on the quantity of interest. We use $S = 1000$ simulated draws to compute the minimum-MSE estimators, since no closed-form solution is available in this case (see Appendix S3).

In Figure 5, we see that the parametric random-effects dynamic probit estimates of δ_{β_0, π_0} and β_0 are substantially biased for $T = 5$ and $T = 10$, whereas the bias is smaller when $T = 20$. By contrast, the minimum-MSE estimator performs better in terms of bias for both quantities of interest, in particular when taking ϵ to be one of the first few ϵ_k's given by (29). In the top panel of Table 2, we show the bias and root MSE of various estimators of δ_{β_0, π_0}: the random-effects estimator based on the normal reference model, an empirical Bayes estimator, the linear probability estimator, and the minimum-MSE estimators based on $\epsilon_1, \epsilon_2, \epsilon_3$.[31] We see that the minimum-MSE estimator dominates

---

[30] While this finding is interesting, note that our CI construction has coverage guarantees only when π_0 belongs to an ϵ-neighborhood of $\pi(\gamma_*)$, which is not the case here since the true distribution of A lies outside the neighborhoods for the range of ϵ that we consider.

[31] The random-effects and empirical Bayes estimators are given by $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\pi(\hat{\gamma})}[\Phi(\hat{\beta} + A) - \Phi(A)]$ and $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\pi(\hat{\gamma})}[\Phi(\hat{\beta} + A) - \Phi(A)|Y = Y_i]$, respectively. In fixed-lengths panels both estimators are consistent

A. Average state dependence δ_{β_0, π_0}

(a) $T = 5$ (b) $T = 10$ (c) $T = 20$

B. Autoregressive parameter β_0

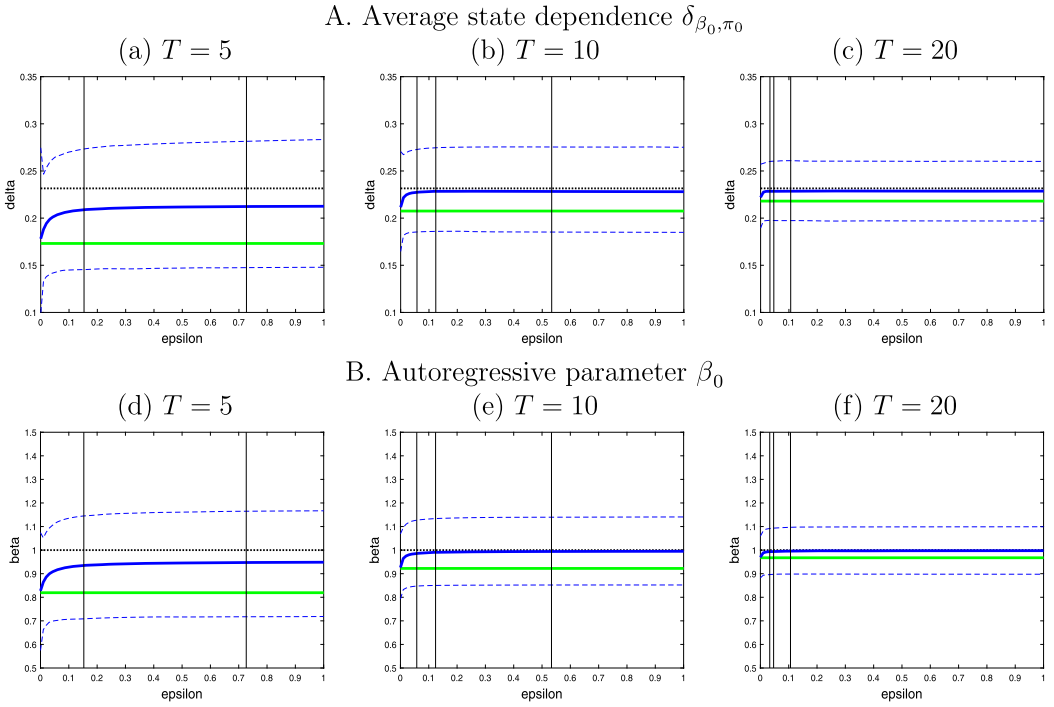(d) $T = 5$ (e) $T = 10$ (f) $T = 20$

FIGURE 5. Minimum-MSE estimator in the dynamic panel binary choice model. *Notes*: The solid horizontal line corresponds to the mean random-effects estimator among 1000 simulations, the solid curve to the mean minimum-MSE estimator (with 2.5% and 97.5% percentiles in dashed), and the dotted horizontal line to the truth. ϵ is reported on the x-axis, and the vertical lines indicate ϵ_k, $k \in \{1, 2\}$ (left column) and $k \in \{1, 2, 3\}$ (middle and right columns). In the left column, ϵ_3 is too large to be included in the figure. $n = 500$.

TABLE 2. Monte Carlo simulation in the dynamic binary choice panel data model: comparison between various estimators.

$T =$	5	10	20	5	10	20
		Bias			Root MSE	
A. Average state dependence δ_{β_0, π_0}						
Random-effects	−0.0585	−0.0252	−0.0140	0.0633	0.0311	0.0198
Empirical Bayes	−0.0574	−0.0215	−0.0053	0.0622	0.0282	0.0141
Linear probability	−0.2491	−0.0976	0.0012	0.2497	0.0990	0.0128
Minimum-MSE (ϵ_1)	−0.0227	−0.0057	−0.0029	0.0397	0.0232	0.0154
Minimum-MSE (ϵ_2)	−0.0194	−0.0048	−0.0028	0.0388	0.0233	0.0154
Minimum-MSE (ϵ_3)	−0.0196	−0.0049	−0.0027	0.0412	0.0235	0.0155
B. Autoregressive parameter β_0						
Maximum likelihood	−0.1804	−0.0817	−0.0328	0.2003	0.1001	0.0506
Minimum-MSE (ϵ_1)	−0.0646	−0.0198	−0.0055	0.1288	0.0747	0.0479
Minimum-MSE (ϵ_2)	−0.0522	−0.0155	−0.0045	0.1258	0.0746	0.0481
Minimum-MSE (ϵ_3)	−0.0432	−0.0116	−0.0030	0.1282	0.0747	0.0486

Note: Performance of various estimators in the dynamic panel data binary choice model, for different values of T. $n = 500$, results for 1000 simulations.

all other estimators, for these ϵ_k values, when $T = 5$ and $T = 10$. In the bottom panel of Table 2, we show the results for the random-effects MLE and minimum-MSE estimators of β_0. The results are similar to the case of average state dependence. In this DGP, minimum-MSE estimators achieve bias reduction under misspecification, even when T is quite small. Bias reduction comes with some increase in variance, yet the overall MSE is lower for minimum-MSE estimators compared to the MLE. Lastly, in Tables S3 and S4 in the Appendix, we show additional information about the simulation results, for the autoregressive parameter and the average state dependence parameter, respectively.

7. Conclusion

We propose a framework for estimation and inference in the presence of model misspecification. This allows researchers to perform sensitivity analysis for existing estimators, and to construct improved estimators and confidence intervals that are less sensitive to model assumptions. Our approach is based on a minimax mean squared error rule, which consists of a one-step adjustment of the initial estimate. This adjustment is motivated by both robustness and efficiency, and it remains valid when the identification of the "large" model is irregular or point-identification fails. Hence, our approach provides a complement to partial identification methods, when the researcher sees her reference model as a plausible, albeit imperfect, approximation to reality. Given a parametric reference model, implementing our estimators and confidence intervals does not require estimating a larger model. This is an attractive feature in complex models such as dynamic structural models, for which sensitivity analysis methods are needed. Lastly, while our theory applies quite generally, we have provided explicit expressions and described implementation in two specific classes of problems: parametric models, and semiparametric likelihood models with a mixture structure. Generalizing the applicability of the approach to other semiparametric models is an important task for future work.

Appendix: Main results

In this section of the Appendix, we provide the proofs for the main results of Section 2. As in the rest of the paper, we always implicitly assume that all functions of y are measurable, and that all expectations and integrals over y are well-defined.

A.1 *Proof of Theorem* 1

A.1.1 *Notation and assumptions*   In all our applications, Π is either a vector space or an affine space. Let $\overline{\mathcal{T}}$ and $\mathcal{T}$ be the tangent and cotangent spaces of Π at $\pi(\gamma_*)$. Thus, for $\pi_1, \pi_2 \in \Pi$ we have $(\pi_1 - \pi_2) \in \overline{\mathcal{T}}$, and $\mathcal{T}$ is the set of linear maps $u : \overline{\mathcal{T}} \to \mathbb{R}$. For a scalar

---

under the parametric reference specification, and the random-effects estimator is efficient. However, the two estimators are generally biased under misspecification. Bonhomme and Weidner (2021) show that the empirical Bayes estimator has minimum local asymptotic worst-case specification error, albeit in neighborhoods of the reference model where the probit conditional likelihood given individual effects and lagged outcomes may be incorrectly specified.

function $q : \Pi \mapsto \mathbb{R}$, we have $\nabla_\pi q_{\pi(\gamma_*)} \in \mathcal{T}$; that is, the typical element of $\mathcal{T}$ is a gradient. Conversely, for a map to Π, such as $\gamma \mapsto \pi(\gamma)$, we have $\frac{\partial \pi(\gamma_*)}{\partial \gamma_k} \in \overline{\mathcal{T}}$.

For $v \in \overline{\mathcal{T}}$ and $u \in \mathcal{T}$, we use the bracket notation $\langle v, u \rangle \in \mathbb{R}$ to denote the bilinear mapping. Here, we are not assuming a Hilbert space structure, and we only use the bracket notation to combine vectors v and covectors u into a scalar.

Our squared distance measure $d(\pi_0, \pi(\gamma_*))$ on Π induces a norm on the tangent space $\overline{\mathcal{T}}$, namely for $v \in \overline{\mathcal{T}}$,

$$\|v\|^2_{\mathrm{ind}, \gamma_*} = \lim_{\epsilon \to 0} \frac{d(\pi(\gamma_*) + \epsilon^{1/2} v, \pi(\gamma_*))}{\epsilon}.$$

Throughout, we assume that $\dim \beta$ and $\dim \gamma$ are finite. For any finite-dimensional vectors, we use the standard Euclidean norm $\|\cdot\|$, and for any finite-dimensional matrices we use the spectral matrix norm, which we also denote by $\|\cdot\|$. Let $\mathcal{Y}$ denote the support of Y.

ASSUMPTION A1. *We assume that $Y_i \sim i.i.d. f_{\beta_0, \pi_0}$. In addition, we impose the following regularity conditions*:

(i) *We consider $n \to \infty$ and $\epsilon \to 0$ such that $\epsilon n \to c$, for some constant $c \in (0, \infty)$.*

(ii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \|\nabla_\pi \delta_{\beta_0, \pi_0}\|_{\gamma_*} = O(1)$, *and* $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} |\delta_{\beta_0, \pi_0} - \delta_{\beta_0, \pi(\gamma_*)} - \langle \pi_0 - \pi(\gamma_*), \nabla_\pi \delta_{\beta_0, \pi(\gamma_*)}\rangle| = o(\epsilon^{1/2})$.

(iii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \{\int_{\mathcal{Y}} [f^{1/2}_{\beta_0, \pi_0}(y) - f^{1/2}_{\beta_0, \pi(\gamma_*)}(y)]^2 \, dy\}^{1/2} = O(\epsilon^{1/2})$,
$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \|\nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(y)\|^2_{\gamma_*} [f^{1/2}_{\beta_0, \pi_0}(y) - f^{1/2}_{\beta_0, \pi(\gamma_*)}(y)]^2 \, dy = o(1)$,
$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} [f^{1/2}_{\beta_0, \pi_0}(y) - f^{1/2}_{\beta_0, \pi(\gamma_*)}(y) - \langle \pi_0 - \pi(\gamma_*), \nabla_\pi f^{1/2}_{\beta_0, \pi(\gamma_*)}(y)\rangle]^2 \, dy = o(\epsilon)$.

(iv) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \epsilon^{-1/2} \|\pi_0 - \pi(\gamma_*)\|_{\mathrm{ind}, \gamma_*} = 1 + o(1)$. *Furthermore, for $u_\epsilon \in \mathcal{T}$ with $\|u_\epsilon\|_{\gamma_*} = O(1)$ we have*

$$\left| \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \epsilon^{-1/2} \langle \pi_0 - \pi(\gamma_*), u_\epsilon \rangle - \|u_\epsilon\|_{\gamma_*} \right| = o(1).$$

(v) *For some $\nu > 0$, we have* $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \|\nabla_{\beta\gamma} \log f_{\beta_0, \pi(\gamma_*)}(Y)\|^{2+\nu} = O(1)$, *and* $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \|\nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(Y)\|^{2+\nu}_{\gamma_*} = O(1)$. *Furthermore, we assume that* $\|\nabla_{\beta\gamma} \delta_{\beta_0, \pi(\gamma_*)}\| = O(1)$, *and* $\|H^{-1}_{\beta\gamma}\| = O(1)$.

Part (i) of Assumption A1 describes our asymptotic framework, where the assumption $\epsilon n \to c$ is required to ensure that the squared worst-case bias (of order ϵ) and the variance (of order $1/n$) of the estimators for δ_{β_0, π_0} are asymptotically of the same order, so that the MSE provides a meaningful balance between bias and variance asymptotically. Part (ii) requires δ_{β_0, π_0} to be sufficiently smooth in π_0, so that a first-order Taylor expansion provides a good local approximation to δ_{β_0, π_0}.

Part (iii) of Assumption A1 is a smoothness assumption on $f_{\beta_0, \pi_0}(y)$ in π_0. Those conditions may not look intuitive, in particular when π_0 is infinite-dimensional, so we

want to discuss that assumption in some more detail here for the case of the semiparametric mixture models introduced in Section 3.4, where $f_{\beta_0, \pi_0}(y) = \int_{\mathcal{A}} g_{\beta_0}(y|a)\pi_0(a)\, da$. In that case, we have

$$\int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0, \pi_0}(y) - f^{1/2}_{\beta_0, \pi(\gamma_*)}(y) \right]^2 dy = 2H^2(f_{\beta_0, \pi_0}, f_{\beta_0, \pi(\gamma_*)})$$

$$\leq 2D_{\mathrm{KL}}(f_{\beta_0, \pi_0}||f_{\beta_0, \pi(\gamma_*)}) \leq 2D_{\mathrm{KL}}(\pi_0||\pi(\gamma_*)),$$

where the first inequality is the general relation $H^2(f_{\beta_0, \pi_0}, f_{\beta_0, \pi(\gamma_*)}) \leq D_{\mathrm{KL}}(f_{\beta_0, \pi_0}||f_{\beta_0, \pi(\gamma_*)})$ between the squared Hellinger distance H^2 and the Kullback–Leibler divergence D_{KL}, and the second inequality is sometimes called the "chain rule" for the Kullback–Leibler divergence, which can be derived by an application of Jensen's inequality. Since we defined our distance measure $d(\pi_0, \pi(\gamma_*))$ in the semiparametric mixture case to be twice the Kullback–Leibler divergence $2D_{\mathrm{KL}}(\pi_0||\pi(\gamma_*)) = 2\mathbb{E}_{\pi_0} \log[\pi_0(A)/\pi(A|\gamma_*)]$ we find that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\{ \int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0, \pi_0}(y) - f^{1/2}_{\beta_0, \pi(\gamma_*)}(y) \right]^2 dy \right\}^{1/2} \leq \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\{ d(\pi_0, \pi(\gamma_*)) \right\}^{1/2} = \epsilon^{1/2}.$$

Thus, the first condition in Assumption A1(iii) is satisfied for those semiparametric mixture models.

The second condition in Assumption A1(iii) can be justified by imposing that

$$\sup_{y \in \mathcal{Y}} \left\| \nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(y) \right\|^2_{\gamma_*} = O(1),$$

which for the semiparametric mixture model can equivalently be written as

$$\sup_{y \in \mathcal{Y}} \frac{\mathrm{Var}_{\pi(\gamma_*)}\left[g_{\beta_0}(y|A) \right]}{\left[\mathbb{E}_{\pi(\gamma_*)} g_{\beta_0}(y|A) \right]^2} = O(1). \tag{A1}$$

For any standard discrete choice model (as those discussed in Section 6), we have that $\sup_{y \in \mathcal{Y}} \mathrm{Var}_{\pi(\gamma_*)}[g_{\beta_0}(y|A)] < \infty$, and $\inf_{y \in \mathcal{Y}} \mathbb{E}_{\pi(\gamma_*)} g_{\beta_0}(y|A) > 0$, implying that equation (A1) is satisfied. We then have

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left\| \nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(y) \right\|^2_{\gamma_*} \left[f^{1/2}_{\beta_0, \pi_0}(y) - f^{1/2}_{\beta_0, \pi(\gamma_*)}(y) \right]^2 dy$$

$$\leq \underbrace{\left[\sup_{y \in \mathcal{Y}} \left\| \nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(y) \right\|^2_{\gamma_*} \right]}_{=O(1)} \underbrace{\left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0, \pi_0}(y) - f^{1/2}_{\beta_0, \pi(\gamma_*)}(y) \right]^2 dy \right\}}_{\leq \epsilon = o(1)} = o(1).$$

Thus, one way to justify the second condition in Assumption A1(iii) is to argue that equation (A1) holds, which is the case for our illustrations in Section 6. The last condition in Assumption A1(iii) could be broken down analogously for semiparametric mixture models, but it is actually a standard condition of differentiability in quadratic mean that is also regularly imposed when π is infinite-dimensional (see, e.g., equation (5.38) in Van der Vaart (2007)).

Part (iv) of Assumption A1 requires that our distance measure $d(\pi_0, \pi(\gamma_*))$ converges to the associated norm for small values ϵ in a smooth fashion. Finally, part (v) requires invertibility of $H_{\beta\gamma}$ (but invertibility of H_π or $\widetilde{H}_\pi$ are *not* required), uniform boundedness of various derivatives, and of the $(2 + \nu)$-th moment of $\nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(Y)$—which again can be justified by equation (A1), because we then have $\sup_{y \in \mathcal{Y}} \|\nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(y)\|_{\gamma_*}^2 = O(1)$.

For many of the proofs, we only need the regularity conditions in Assumption A1. However, in order to describe the properties of our minimum-MSE estimator $\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \delta_{\widehat{\beta}, \pi(\widehat{\gamma})} + \frac{1}{n} \sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i, \widehat{\beta}, \widehat{\gamma})$ we also need to account for the fact that $\widehat{\beta}$ and $\widehat{\gamma}$ themselves are estimated. It turns our that the leading-order asymptotic properties of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ are independent of whether β_0 and γ_* are known or estimated in the construction of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ (see, e.g., Lemma A3 below), but formally showing this requires some additional assumptions, which we present next.

ASSUMPTION A2. *For some $\chi > 2$, we have*:

(i) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} (\mathbb{E}_{\beta_0, \pi_0} \| \binom{\widehat{\beta}}{\widehat{\gamma}} - \binom{\beta_0}{\gamma_*} \|^\chi)^{\frac{1}{\chi}} = O(\frac{1}{\sqrt{n}})$.

(ii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \|\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y, \beta_0, \gamma_*)\| = O(1)$, *where* $\eta = (\beta', \gamma')'$.

(iii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \sup_{\beta \in \mathcal{B}, \gamma \in \mathcal{G}} \| \frac{1}{n} \sum_{i=1}^n \nabla_{\eta\eta'}^2 h_\epsilon^{\mathrm{MMSE}}(Y_i, \beta, \gamma)\| = O(1)$, *where* $\eta = (\beta', \gamma')'$.

(iv) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} [h_\epsilon^{\mathrm{MMSE}}(Y, \beta_0, \gamma_*)]^{2+\nu} = O(1)$, *for some $\nu > 0$*.

Part (i) of Assumption A2 requires $\widehat{\beta}$ and $\widehat{\gamma}$ to converge at $\sqrt{n}$ rate. As discussed in the main text, we assume that preliminary estimators have finite χ-moments where $\chi > 2$. Part (ii) of Assumption A2 requires a uniformly bounded second moment for $\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(y, \beta_0, \gamma_*)$. Since equation (20) in the main text gives an explicit expression for $h_\epsilon^{\mathrm{MMSE}}(y, \beta_0, \gamma_*)$, we could replace Assumption A2(ii) by appropriate assumptions on the model primitives $f_{\beta_0, \pi_0}(y)$ and δ_{β_0, π_0}, but for the sake of brevity we state the assumption in terms of $h_\epsilon^{\mathrm{MMSE}}(y, \beta_0, \gamma_*)$. The same is true for part (iii) of Assumption A2. Notice that this last part of the assumption involves a supremum over β and γ inside of an expectation—in order to verify it, one either requires a uniform Lipschitz bound on the dependence of $h_\epsilon^{\mathrm{MMSE}}(Y_i, \beta, \gamma)$ on β and γ, or some empirical process method to control the entropy of that function (e.g., a bracketing argument). But since β and γ are finite-dimensional parameters these are all standard arguments.

We verified Assumption A2(iv) in the locally quadratic case of Section 3. Formally, we have the following lemma.

LEMMA A1. *Let Assumption A1 hold, and assume that $h_\epsilon^{\mathrm{MMSE}}(\cdot, \beta_0, \gamma_*)$ is given by Lemma 1 in the main text. Then Assumption A2(iv) holds with the constant ν specified in Assumption A1.*

A.1.2 *Proof of Theorem* 1   For a function $h_\epsilon = h_\epsilon(y, \beta_0, \gamma_*)$, we define

$$\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) := \delta_{\beta_0, \pi(\gamma_*)} + \frac{1}{n} \sum_{i=1}^n h_\epsilon(Y_i, \beta_0, \gamma_*).$$

It is useful to establish some preliminary lemmas before showing the main result. The proofs for those lemmas are provided in Section S1.1.

LEMMA A2. *Let Assumption* A1 *hold, and let* $h_\epsilon(\cdot, \beta_0, \gamma_*)$ *be a sequence of influence functions that satisfy the unbiasedness constraint* (2) *as well as* $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} |h_\epsilon(Y, \beta_0, \gamma_*)|^\kappa = O(1)$, *for some* $\kappa > 2$. *Then*

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \big[\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0} \big]^2$$

$$= b_\epsilon(h_\epsilon, \beta_0, \gamma_*)^2 + \frac{\mathrm{Var}_{\beta_0, \pi(\gamma_*)}\big(h_\epsilon(Y, \beta_0, \gamma_*)\big)}{n} + o(\epsilon).$$

Lemma A2 provides a formal justification for the worst-case MSE approximation introduced in equation (9) of the main text.

Recall that $\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \widehat{\delta}(h_\epsilon^{\mathrm{MMSE}}, \widehat{\beta}, \widehat{\gamma})$. This differs from $\widehat{\delta}(h_\epsilon^{\mathrm{MMSE}}, \beta_0, \gamma_*)$, because β_0 and γ_* have to be estimated. The following lemma shows that the fact that β_0 and γ_* are estimated in the construction of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ can be neglected to first order. Notice that this result requires the additional regularity conditions in Assumption A2, which are not required anywhere else in the proof of Theorem 1.

LEMMA A3. *Let Assumptions* A1 *and* A2 *hold. Then*

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \big| \widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \widehat{\delta}(h_\epsilon^{\mathrm{MMSE}}, \beta_0, \gamma_*) \big| = O\Big(\frac{1}{n}\Big).$$

Thus, Lemma A3 guarantees that $\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \widehat{\delta}(h_\epsilon^{\mathrm{MMSE}}, \beta_0, \gamma_*) + O_{P_0}(1/n)$. This may be surprising given that the differences $\widehat{\beta} - \beta_0$ and $\widehat{\gamma} - \gamma_*$ are themselves of order $1/\sqrt{n}$. However, recall that by construction $h_\epsilon^{\mathrm{MMSE}}$ satisfies the local robustness condition (3), which is imposed through our constraints (2) and (4). Local robustness ensures that $\widehat{\beta} - \beta_0$ and $\widehat{\gamma} - \gamma_*$ have no leading-order effect on $\widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \widehat{\delta}(h_\epsilon^{\mathrm{MMSE}}, \beta_0, \gamma_*)$.

For the next lemma, recall the decomposition of $\widehat{\delta}_\epsilon$ in Theorem 1 in the main text:

$$\widehat{\delta}_\epsilon = \delta_{\beta_0, \pi(\gamma_*)} + \frac{1}{n} \sum_{i=1}^n h_\epsilon(Y_i, \beta_0, \gamma_*) + n^{-1/2} R_n$$

$$= \widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) + n^{-1/2} R_n. \tag{A2}$$

Here, $\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*)$ is the well-behaved leading-order contribution to $\widehat{\delta}_\epsilon$, whereas R_n is an asymptotically vanishing remainder term that may, however, have heavy tails (it only satisfies a trimmed second moment condition). The following lemma shows that the worst-case trimmed MSE of $\widehat{\delta}_\epsilon$ is bounded from below by the MSE of the leading-order term $\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*)$.

LEMMA A4. *Let Assumption* A1 *hold, and let* $h_\epsilon(\cdot, \beta_0, \gamma_*)$ *be a sequence of influence functions that satisfy the unbiasedness constraint* (2) *as well as* $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} |h_\epsilon(Y, \beta_0, \gamma_*)|^\kappa = O(1)$, *for some* $\kappa > 2$. *Assume that* (A2) *holds, and let* $m_n > 0$ *be a sequence such that* $m_n n^{1/2} [\log(n)]^{-1} \to \infty$. *Furthermore, assume that:*

(i) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} P_{\beta_0, \pi_0}(|R_n| > \log(n)) = o(1)$,

(ii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}[R_n^2 \mathbb{1}(|R_n| \le 2\log(n))] = o(1)$.

Then we have

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0}\big)^2\big]$$

$$\le \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[(\widehat{\delta}_\epsilon - \delta_{\beta_0, \pi_0})^2 \mathbb{1}(|\widehat{\delta}_\epsilon - \delta_{\beta_0, \pi_0}| \le m_n)\big] + o(\epsilon). \tag{A3}$$

We now have all the preliminary results required to show the main theorem.

PROOF OF THEOREM 1.   Define

$$r_\epsilon := \widehat{\delta}_\epsilon^{\text{MMSE}} - \widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*).$$

We then have

$$\mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}\big)^2 \mathbb{1}(|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}| \le m_n)\big]$$

$$= \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0} + r_\epsilon\big)^2 \mathbb{1}(|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}| \le m_n)\big]$$

$$= \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0}\big)^2 \underbrace{\mathbb{1}(|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}| \le m_n)}_{\le 1}\big]$$

$$+ 2\mathbb{E}_{\beta_0, \pi_0}\big[r_\epsilon\big(\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0} + r_\epsilon\big)\mathbb{1}(|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}| \le m_n)\big]$$

$$\underbrace{- \mathbb{E}_{\beta_0, \pi_0}\big[r_\epsilon^2 \mathbb{1}(|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}| \le m_n)\big]}_{\le 0}$$

$$\le \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0}\big)^2\big]$$

$$+ 2\mathbb{E}_{\beta_0, \pi_0}\big[\underbrace{r_\epsilon\big(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}\big)\mathbb{1}(|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}| \le m_n)}_{\le |r_\epsilon| m_n}\big]$$

$$\le \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0}\big)^2\big] + 2m_n \mathbb{E}_{\beta_0, \pi_0}|r_\epsilon|.$$

According to Lemma A3 we have $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}|r_\epsilon| = O(1/n) = O(\epsilon)$, and the assumptions of the theorem guarantee that $m_n = o(1)$. We thus obtain

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}\big)^2 \mathbb{1}(|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0, \pi_0}| \le m_n)\big]$$

$$\le \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0}\big)^2\big] + o(\epsilon). \tag{A4}$$

By definition, $h_\epsilon^{\mathrm{MMSE}}$ also satisfies the unbiasedness constraint (2). Together with Assumption A2(iv) this implies that $h_\epsilon^{\mathrm{MMSE}}$ satisfies the conditions on h_ϵ in Lemma A2 with $\kappa = 2 + \nu$. Thus, we can apply Lemma A2 with $h_\epsilon = h_\epsilon^{\mathrm{MMSE}}$ to find that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}\big(h_\epsilon^{\mathrm{MMSE}}, \beta_0, \gamma_*\big) - \delta_{\beta_0, \pi_0}\big)^2\big]$$

$$= b_\epsilon\big(h_\epsilon^{\mathrm{MMSE}}, \beta_0, \gamma_*\big)^2 + \frac{\mathrm{Var}_{\beta_0, \pi(\gamma_*)}\big(h_\epsilon^{\mathrm{MMSE}}(Y, \beta_0, \gamma_*)\big)}{n} + o(\epsilon). \tag{A5}$$

The function $h_\epsilon^{\mathrm{MMSE}}(\cdot, \beta_0, \gamma_*)$ is defined by the minimization problem (10) in the main text. In other words, $h_\epsilon^{\mathrm{MMSE}}(\cdot, \beta_0, \gamma_*)$ minimizes the objective function $b_\epsilon(h, \beta_0, \gamma_*)^2 + n^{-1}\mathrm{Var}_{\beta_0, \pi(\gamma_*)}(h(Y, \beta_0, \gamma_*))$, subject to the constraints (2) and (4). Theorem 1 assumes that $h_\epsilon = h_\epsilon(\cdot, \beta_0, \gamma_*)$ satisfies those constraints, and the definition of $h_\epsilon^{\mathrm{MMSE}}(\cdot, \beta_0, \gamma_*)$ therefore implies that

$$b_\epsilon\big(h_\epsilon^{\mathrm{MMSE}}, \beta_0, \gamma_*\big)^2 + \frac{\mathrm{Var}_{\beta_0, \pi(\gamma_*)}\big(h_\epsilon^{\mathrm{MMSE}}(Y, \beta_0, \gamma_*)\big)}{n}$$

$$\leq b_\epsilon(h_\epsilon, \beta_0, \gamma_*)^2 + \frac{\mathrm{Var}_{\beta_0, \pi(\gamma_*)}\big(h_\epsilon(Y, \beta_0, \gamma_*)\big)}{n}. \tag{A6}$$

Theorem 1 also imposes all the assumptions on h_ϵ in Lemma A2. By applying that lemma, we thus have

$$b_\epsilon(h_\epsilon, \beta_0, \gamma_*)^2 + \frac{\mathrm{Var}_{\beta_0, \pi(\gamma_*)}\big(h_\epsilon(Y, \beta_0, \gamma_*)\big)}{n}$$

$$= \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0, \pi_0}\big)^2\big] + o(\epsilon). \tag{A7}$$

Finally, Theorem 1 also guarantees all the assumptions of Lemma A4, implying that the inequality (A3) holds. Now, combining (A4), (A5), (A6), (A7), and (A3) gives

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[\big(\widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \delta_{\beta_0, \pi_0}\big)^2 \mathbb{1}\big(\big|\widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \delta_{\beta_0, \pi_0}\big| \leq m_n\big)\big]$$

$$\leq \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0}\big[(\widehat{\delta}_\epsilon - \delta_{\beta_0, \pi_0})^2 \mathbb{1}\big(|\widehat{\delta}_\epsilon - \delta_{\beta_0, \pi_0}| \leq m_n\big)\big] + o(\epsilon), \tag{A8}$$

which is what we wanted to show.   □

A.2 *Proof of Theorem 2*

ASSUMPTION A3.

(i) $\widehat{\delta} - \delta_{\beta_0, \pi(\gamma_*)} - \frac{1}{n}\sum_{i=1}^n h(Y_i, \beta_0, \gamma_*) = o_{P_{\beta_0, \pi_0}}(n^{-\frac{1}{2}})$, *uniformly in* $\pi_0 \in \Gamma_\epsilon(\gamma_*)$.

(ii) *Let* $\sigma_h^2(\beta_0, \pi_0, \gamma_*) = \mathrm{Var}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*)$. *We assume that there exists a constant* c, *independent of* ϵ, *such that* $\inf_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \sigma_h(\beta_0, \pi_0, \gamma_*) \geq c > 0$. *Furthermore, for*

all sequences $a_n = c_{1-\alpha/2} + o(1)$ we have

$$\inf_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \Pr_{\beta_0, \pi_0} \left[\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \right| \le a_n \right] \ge 1 - \alpha + o(1).$$

(iii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \| \widehat{\beta} - \beta_0 \|^2 = o(1)$, $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \| \widehat{\gamma} - \gamma_* \|^2 = o(1)$, $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} [\widehat{\sigma}_h - \sigma_h(\beta_0, \pi_0, \gamma_*)]^2 = o(1)$.

(iv) $\| \nabla_{\beta\gamma} b_\epsilon(h, \beta, \gamma) \| = O(\epsilon^{\frac{1}{2}})$, *uniformly in some neighborhood around* β_0, γ_*.

Part (i) is weaker than the local regularity of the estimator $\widehat{\delta}$ that we assumed when analyzing the minimum-MSE estimator; see equation (14). In turn, related to but differently from the conditions we used for Theorem 1, part (ii) requires a form of local asymptotic normality of the estimator.

PROOF OF THEOREM 2. Let $\widehat{\delta}$ be an estimator and $h(y, \beta_0, \gamma_*)$ be the corresponding influence function such that part (i) in Assumption A3 holds. Define $\widehat{R}_{\beta_0, \gamma_*} := \widehat{\delta} - \delta_{\beta_0, \pi(\gamma_*)} - \frac{1}{n} \sum_{i=1}^{n} h(Y_i, \beta_0, \gamma_*)$. We then have

$$\widehat{\delta} - \delta_{\beta_0, \pi_0} = \frac{1}{n} \sum_{i=1}^{n} h(Y_i, \beta_0, \gamma_*) + \delta_{\beta_0, \pi(\gamma_*)} - \delta_{\beta_0, \pi_0} + \widehat{R}_{\beta_0, \gamma_*}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*) \right]$$

$$- \left[\delta_{\beta_0, \pi_0} - \delta_{\beta_0, \pi(\gamma_*)} - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*) \right] + \widehat{R}_{\beta_0, \gamma_*},$$

and, therefore,

$$\underbrace{\frac{|\widehat{\delta} - \delta_{\beta_0, \pi_0}| - b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) - \widehat{\sigma}_h c_{1-\alpha/2}/\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)/\sqrt{n}}}_{=\text{lhs}}$$

$$\le \underbrace{\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \right| - c_{1-\alpha/2} + \widehat{r}_{\beta_0, \pi_0, \gamma_*}}_{=\text{rhs}}, \qquad (A9)$$

where

$$\widehat{r}_{\beta_0, \pi_0, \gamma_*}$$

$$:= c_{1-\alpha/2}$$

$$+ \frac{\left| \delta_{\beta_0, \pi_0} - \delta_{\beta_0, \pi(\gamma_*)} - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*) \right| + |\widehat{R}_{\beta_0, \gamma_*}| - b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) - \widehat{\sigma}_h c_{1-\alpha/2}/\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)/\sqrt{n}}$$

$$= \frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \left\{ \left| \delta_{\beta_0, \pi_0} - \delta_{\beta_0, \pi(\gamma_*)} - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*) \right| + |\widehat{R}_{\beta_0, \gamma_*}| \right.$$

$$\left. - b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) - \frac{\widehat{\sigma}_h - \sigma_h(\beta_0, \pi_0, \gamma_*)}{\sqrt{n}} c_{1-\alpha/2} \right\}.$$

From (A9), we conclude that the event rhs ≤ 0 implies the event lhs ≤ 0 and, therefore, $\Pr_{\beta_0, \pi_0}(\text{lhs} \leq 0) \geq \Pr_{\beta_0, \pi_0}(\text{rhs} \leq 0)$, which we can also write as

$$\Pr_{\beta_0, \pi_0}\left[|\widehat{\delta} - \delta_{\beta_0, \pi_0}| \leq b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) + \frac{\widehat{\sigma}_h}{\sqrt{n}} c_{1-\alpha/2} \right]$$

$$\geq \Pr_{\beta_0, \pi_0}\left[\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \right| \leq c_{1-\alpha/2} - \widehat{r}_{\beta_0, \pi_0, \gamma_*} \right]. \quad (A10)$$

By part (iv) in Assumption A3 there exists a constant $C > 0$ such that $\|\nabla_{\beta\gamma} b_\epsilon(h, \beta, \gamma)\| \leq C\epsilon^{\frac{1}{2}}$, uniformly in a neighborhood of (β_0, γ_*) and, therefore,

$$\left| b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) - b_\epsilon(h, \beta_0, \gamma_*) \right| \leq C\epsilon^{\frac{1}{2}} \left\| \begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\gamma} - \gamma_* \end{pmatrix} \right\|.$$

Using this, we find that

$$|\widehat{r}_{\beta_0, \pi_0, \gamma_*}| \leq \frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \left\{ \left| |\delta_{\beta_0, \pi_0} - \delta_{\beta_0, \pi(\gamma_*)} - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*)| - b_\epsilon(h, \beta_0, \gamma_*) \right| \right.$$

$$\left. + \frac{|\widehat{\sigma}_h - \sigma_h(\beta_0, \pi_0, \gamma_*)|}{\sqrt{n}} c_{1-\alpha/2} + C\epsilon^{\frac{1}{2}} \left\| \begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\gamma} - \gamma_* \end{pmatrix} \right\| + |\widehat{R}_{\beta_0, \gamma_*}| \right\}.$$

Parts (i) and (ii) of Assumption A3 imply that, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, we have

$$\frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \widehat{R}_{\beta_0, \gamma_*} = o_{P_{\beta_0, \pi_0}}(1),$$

and analogously we find from the conditions in Assumption A3 that

$$\frac{\widehat{\sigma}_h - \sigma_h(\beta_0, \pi_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)} = o_{P_{\beta_0, \pi_0}}(1), \quad \frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \epsilon^{\frac{1}{2}} \left\| \begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\gamma} - \gamma_* \end{pmatrix} \right\| = o_{P_{\beta_0, \pi_0}}(1),$$

uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Finally, since we also impose Assumption A1 and $\sup_{\pi_0 \in \Gamma_\epsilon} \mathbb{E}_{\beta_0, \pi_0} h^2(Y, \beta_0, \gamma_*) = O(1)$ we obtain, analogously to the proof of Lemma S1(iii) in Section S1, that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \left| |\delta_{\beta_0, \pi_0} - \delta_{\beta_0, \pi(\gamma_*)} - \mathbb{E}_{\beta_0, \pi_0} h(Y, \beta_0, \gamma_*)| - b_\epsilon(h, \beta_0, \gamma_*) \right| = o(1).$$

We thus conclude that $\widehat{r}_{\beta_0, \pi_0, \gamma_*} = o_{P_{\beta_0, \pi_0}}(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Together with (A10) and part (ii) in Assumption A3 this implies (17), hence Theorem 2. $\qquad\square$

REFERENCES

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005), "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools." *Journal of Political Economy*, 113, 151–184. [910]

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro (2017), "Measuring the sensitivity of parameter estimates to estimation moments." *Quarterly Journal of Economics*, 132 (4), 1553–1592. [910]

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro (2020), "On the informativeness of descriptive statistics for structural estimates." *Econometrica*, 88 (6), 2231–2258. [910]

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters (2017), "Leveraging lotteries for school value-added: Testing and estimation." *Quarterly Journal of Economics*, 132 (2), 871–919. [910]

Armstrong, Timothy B. and Michal Kolesár (2020), "Simple and honest confidence intervals in nonparametric regression." *Quantitative Economics*, 11 (1), 1–39. [908, 918]

Armstrong, Timothy B. and Michal Kolesár (2021), "Sensitivity analysis using approximate moment condition models." *Quantitative Economics*, 12 (1), 77–108. [910, 918, 929]

Attanasio, Orazio P., Costas Meghir, and Ana Santiago (2012), "Education choices in Mexico: Using a structural model and a randomized experiment to evaluate progresa." *The Review of Economic Studies*, 79 (1), 37–66. [909, 929]

Barden, Dennis and Charles B. Thomas (2003), *An Introduction to Differential Manifolds*. World Scientific. [919]

Berger, James and L. Mark Berliner (1986), "Robust Bayes and empirical Bayes analysis with ε-contaminated priors." Annals of Statistics, 461–486. [910]

Bickel, Peter J., Christopher A. J. Klaassen, Yaatov Ritov, and J. A. Wellner (1993), *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press. [913, 916]

Bonhomme, Stéphane (2012), "Functional differencing." *Econometrica*, 80 (4), 1337–1385. [925]

Bonhomme, Stéphane and Laurent Davezies (2017), "Panel data, inverse problems, and the estimation of policy parameters." unpublished manuscript. [925]

Bonhomme, Stéphane and Martin Weidner (2018), "Minimizing sensitivity to model misspecification." cemmap working paper (CWP59/18). [912]

Bonhomme, Stéphane and Martin Weidner (2021), "Posterior average effects." the Journal of Business and Economic Statistics. to appear in. [940]

Bonhomme, Stéphane, Martin Weidner (2022), "Supplement to 'Minimizing sensitivity to model misspecification'." *Quantitative Economics Supplemental Material*, 13, https://doi.org/10.3982/QE1930. [911]

Bugni, Federico A., Ivan A. Canay, and Patrik Guggenberger (2012), "Distortions of asymptotic confidence size in locally misspecified moment inequality models." *Econometrica*, 80 (4), 1741–1768. [910]

Bugni, Federico A. and Tetsuya Ura (2019), "Inference in dynamic discrete choice problems under local misspecification." *Quantitative Economics*, 10 (1), 67–103. [910]

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson (2014), "Small bandwidth asymptotics for density-weighted average derivatives." *Econometric Theory*, 30 (1), 176–200. [913]

Chamberlain, Gary (1984), "Panel data." In *Handbook of Econometrics*, Vol. 2 (Z. Griliches and M. D. Intriligator, eds.). Elsevier Science, Amsterdam. [923]

Chamberlain, Gary (2000), "Econometrics and decision theory." *Journal of Econometrics*, 95 (2), 255–283. [910]

Chamberlain, Gary (2010), "Binary response models for panel data: Identification and information." *Econometrica*, 78 (1), 159–168. [909, 925, 938]

Chen, Xiaohong, Elie T. Tamer, and Alexander Torgovitsky (2011), "Sensitivity analysis in semiparametric likelihood models." unpublished manuscript. [910]

Cheng, Xiaohong and Zhipeng Liao (2015), "Select the valid and relevant moments: An information-based LASSO for GMM with many moments." *Journal of Econometrics*, 186 (2), 443–464. [913]

Chernozhukov, Victor, Dennis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James J. Robins (2018), "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, 21 (1), C1–C68. [909]

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, and Whitney K. Newey (2020), "Locally robust semiparametric estimation." arXiv:1608.00033. Preprint. [910, 913]

Chernozhukov, Victor, Ivan Fernández-Val, Jinyong Hahn, and Whitney Newey (2013), "Average and quantile effects in nonseparable panel models." *Econometrica*, 81 (2), 535–580. [909, 925]

Christensen, Timothy and Benjamin Connault (2019), "Counterfactual sensitivity and robustness." arXiv:1904.00989v2. [910, 927]

Claeskens, Gerda and Niels L. Hjort (2003), "The focused information criterion." *Journal of the American Statistical Association*, 98 (464), 900–916. [910]

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi (2012), "Plausibly exogenous." *Review of Economics and Statistics*, 94 (1), 260–272. [910, 926]

Donoho, Don L. (1994), "Statistical estimation and optimal recovery." The Annals of Statistics, 238–270. [908, 918]

Engl, Hans W., Martin Hanke, and Andreas Neubauer (2000), *Regularization of Inverse Problems*. Kluwer Academic Publishers. [926]

Fessler, Pirmin and Maximilian Kasy (2019), "How to use economic theory to improve estimators." *Review of Economics and Statistics*, 101 (4), 681–698. [910]

Fraser, Donald A. S. (1964), "On local unbiased estimation." Journal of the Royal Statistical Society Series B (Methodological), 46–51. [910]

Giacomini, Raffaella and Toru Kitagawa (2021), "Robust Bayesian inference for set-identified models." *Econometrica*, 89 (4), 1519–1556. [910]

Guggenberger, Patrik (2012), "On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption." *Econometric Theory*, 28 (2), 387–421. [910]

Gustafson, Paul (2000), "Local robustness in Bayesian analysis." In *Robust Bayesian Analysis*, 71–88, Springer, New York, NY. [910]

Hahn, Jinyong and Jerry Hausman (2005), "Estimation with valid and invalid instruments." Annales d'Economie et de Statistique, 25–57. [910]

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel (1986), *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. [909]

Hansen, Bruce E. (2016), "Efficient shrinkage in parametric models." *Journal of Econometrics*, 190(1), 115–132. [910]

Hansen, Bruce E. (2017), "Stein-like 2SLS estimator." *Econometric Reviews*, 36 (6-9), 840–852. [910]

Hansen, Lars P. and M. Marinacci (2016), "Ambiguity aversion and model misspecification: An economic perspective." *Statistical Science*, 31 (4), 511–515. [910]

Hansen, Lars P. and T. J. Sargent (2001), "Robust control and model uncertainty." *American Economic Review*, 91 (2), 60–66. [908]

Hansen, Lars P. and T. J. Sargent (2008), *Robustness*. Princeton University Press. [908, 910, 927]

Hausman, Jerry, Randall Lewis, Konrad Menzel, and Whitney Newey (2011), "Properties of the CUE estimator and a modification with moments." *Journal of Econometrics*, 165 (1), 45–57. [918]

Heckman, James J. (1981), "An incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process." In *The Structural Analysis of Discrete Data*, 179–195.

Honoré, Bo E. and Elie Tamer (2006), "Bounds on parameters in panel dynamic discrete choice models." *Econometrica*, 74 (3), 611–629. [909, 925, 938]

Huber, Peter J. (1964), "Robust estimation of a location parameter." *Annals of Mathematical Statistics*, 35 (1), 73–101. [909]

Huber, Peter J. and Elvezio M. Ronchetti (2009), *Robust Statistics*, Second Edition edition. Wiley. [909, 927]

Imbens, Guido (2003), "Sensitivity to exogeneity assumptions in program evaluation." *American Economic Review*, 93, 126–132. [910]

Kitamura, Yuichi, Taisuke Otsu, and Kirill Evdokimov (2013), "Robustness, infinitesimal neighborhoods, and moment restrictions." *Econometrica*, 81 (3), 1185–1201. [910]

Klaassen, Chris A. (1987), "Consistent estimation of the influence function of locally asymptotically linear estimators." The Annals of Statistics, 1548–1562. [913]

Kress, Rainer (2014), *Linear Integral Equations*. Applied Mathematical Sciences., Vol. 82. Springer, New York. [926]

Leamer, Edward (1983), "Let's take the con out of econometrics." *American Economic Review*, 73 (1), 31–43. [928]

Leamer, Edward (1985), "Sensitivity analyses would help." *American Economic Review*, 75 (3), 308–313. [910, 928]

Liao, Zhipeng (2013), "Adaptive GMM shrinkage estimation with consistent moment selection." *Econometric Theory*, 29 (5), 857–904. [913]

Maasoumi, Esfandiar (1978), "A modified Stein-like estimator for the reduced form coefficients of simultaneous equations." Econometrica, 695–703. [910]

Manski, Charles F. (1988), "Identification of binary response models." *Journal of the American statistical Association*, 83 (403), 729–738. [909, 935]

Masten, Matthew A. and Alexandre Poirier (2020), "Inference on breakdown frontiers." *Quantitative Economics*, 11 (1), 41–111. [910]

Masten, Matthew A. and Alexandre Poirier (2021), "Salvaging falsified instrumental variable models." *Econometrica*, 89 (3), 1449–1469. [910]

Mueller, Ulrich K. (2012), "Measuring prior sensitivity and prior informativeness in large Bayesian models." *Journal of Monetary Economics*, 59 (6), 581–597. [910]

Mukhin, Yaroslav (2018), "Sensitivity of regular estimators." arXiv:1805.08883. arXiv preprint. [910]

Nevo, Aviv and Adam M. Rosen (2012), "Identification with imperfect instruments." *Review of Economics and Statistics*, 94 (3), 659–671. [910]

Newey, W. K. (1994), "The asymptotic variance of semiparametric estimators." *Econometrica*, 6, 1349–1382. [909]

Newey, Whitney K. (1985), "Generalized method of moments specification testing." *Journal of Econometrics*, 29 (3), 229–256. [910]

Newey, Whitney K. (1990), "Semiparametric efficiency bounds." *Journal of Applied Econometrics*, 5 (2), 99–135. [913]

Neyman, Jerzy (1959), "Optimal asymptotic tests of composite statistical hypotheses." In *Probability and Statistics, the Harald Cramer Volume* (U. Grenander, ed.). New York, Wiley. [909]

Norets, Andriy and Xun Tang (2014), "Semiparametric inference in dynamic binary choice models." *Review of Economic Studies*, 81 (3), 1229–1262. [910]

Oster, Emily (2019), "Unobservable selection and coefficient stability: Theory and evidence." *Journal of Business & Economic Statistics*, 37 (2), 187–204. [910]

Pakes, Ariel and Jack Porter (2013), "Moment inequalities for semiparametric multinomial choice with fixed effects." unpublished manuscript. [925]

Rieder, Helmut (1994), *Robust Asymptotic Statistics*. Springer Verlag, New York, NY. [909, 911]

Rosenbaum, Paul R. and Donald B. Rubin (1983), "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *Journal of the Royal Statistical Society Series B*, 45, 212–218. [910]

Schennach, Susanne M. (2014), "Entropic latent variable integration via simulation." *Econometrica*, 82 (1), 345–385. [910]

Severini, Thomas A. and Gautam Tripathi (2012), "Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors." *Journal of Econometrics*, 170 (2), 491–498. [925]

Todd, Petra E. and Kenneth I. Wolpin (2006), "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility." *American Economic Review*, 96 (5), 1384–1417. [909, 929, 932, 933]

Todd, Petra E. and Kenneth I. Wolpin (2008), "Ex ante evaluation of social programs." Annales d'Economie et de Statistique, 263–291. [909, 933]

Van der Vaart, Aad W. (2007), *Asymptotic Statistics*. Cambridge University Press. [912, 927, 942]

Vidakovic, Brani (2000), "Γ-minimax: A paradigm for conservative robust Bayesians." In *Robust Bayesian Analysis*, 241–259, Springer, New York, NY. [910]

Wald, Abraham (1950), *Statistical Decision Functions*. Wiley, New York. [910]

Watson, James and Chris Holmes (2016), "Approximate models and robust decisions." *Statistical Science*, 31 (4), 465–489. [910]

Wolpin, Kenneth I. (2013), *The Limits of Inference Without Theory*. MIT Press. [909, 929, 930]