cemmap

centre for microdata methods and practice

Minimizing Sensitivity to Model Misspecification

Stéphane Bonhomme
Martin Weidner

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper
CWP37/20

Economic
and Social
Research Council

Minimizing Sensitivity to Model Misspecification*

Stéphane Bonhomme† Martin Weidner‡

Revised Draft: July 3, 2020

Abstract

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 simple one-step adjustments. In addition, we provide confidence intervals that contain the true parameter under local misspecification. 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 two binary choice models: a cross-sectional model where the error distribution is misspecified, and a dynamic panel data model where the number of time periods is small and the distribution of individual effects is misspecified.

JEL codes: C13, C23.
Keywords: Model misspecification, robustness, sensitivity analysis, latent variables.

*We thank Kirill Evdokimov, Josh Angrist, Tim Armstrong, Gary Chamberlain, Tim Christensen, Ben Connault, Max Farrell, Jin Hahn, Chris Hansen, Lars Hansen, Kei Hirano, Max Kasy, Roger Koenker, Thibaut Lamadon, Esfandiar Maasoumi, Magne Mogstad, Roger Moon, Whitney Newey, Tai Otsu, Franco Peracchi, Jack Porter, Andres Santos, Azeem Shaikh, Jesse Shapiro, Richard Smith, Alex Torgovistky, and Ken Wolpin, as well as the audiences in various seminars and conferences, for comments. Bonhomme acknowledges support from the NSF, Grant SES-1658920. Weidner acknowledges support from the Economic and Social Research Council through grants (RES-589-28-0001, RES-589-28-0002 and ES/P008909/1) to the ESRC Centre for Microdata Methods and Practice (CeMMAP), and from the European Research Council grants ERC-2014-CoG-646917-ROMIA and ERC-2018-CoG-819086-PANEDA.

†University of Chicago.
‡University College London.

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 specification 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 of the model-based estimates, which do not require re-estimating 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. 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 semi-parametric likelihood 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. To help interpretation, we show that the neighborhood size can be mapped to the local power 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 pre-specified 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, 2016).

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 consider two examples as main illustrations. We first 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 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. We also use this example to illustrate the interpretation of the neighborhood size.

Our second example is a dynamic binary choice model in a short panel data. We assume that time-varying errors are *i.i.d.* normal, but leave unrestricted the distribution of individual heterogeneity given initial conditions. In this setting also, common parameters and average effects often fail to be point-identified (Chamberlain, 2010, Honoré and Tamer, 2006, Chernozhukov *et al.*, 2013), thus motivating a sensitivity analysis approach. We show that minimizing the 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 semi-parametric 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.* (2016), Chernozhukov *et al.* (2018), 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, 2014, Masten and Poirier, 2017). 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 (2018). 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.* (2018) 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 (2018) 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, 2013, Giacomini and Kitagawa, 2018). Christensen and Connault (2018) 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, 2018, 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 semi-parametric models. In Section 4 we discuss the interpretation of neighborhood size. In Section 5 we report simulation exercises in binary choice models, and we conclude in Section 6.

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 the locally-quadratic case, which includes both parametric misspecification and semi-parametric misspecification of distributional functional forms.

2.1 Setup

We observe a random sample $(Y_i : i = 1, \dots, n)$ from the distribution $f_{\beta,\pi}(y) = f(y \,|\, \beta, \pi)$, 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, π_0 that generate the observed data $Y_1, \dots, Y_n$ are unknown to the researcher. Our goal is to estimate δ_{β_0,π_0} and construct confidence intervals around it.

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. Given a distance measure d on Π we denote the maximal amount of misspecification by $\epsilon \geq 0$; that is, we assume that there exists some parameter value $\gamma_* \in \mathcal{G}$ such that the true π_0 satisfies $d(\pi_0, \pi(\gamma_*)) \leq \epsilon$.

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 $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. We will refer to γ_*

as a "pseudo-true" parameter, and 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.

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 Subsection 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 semi-parametric 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 semi-parametric neighborhoods, and we take $d(\pi_0, \pi) = 2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right)\pi_0(a)da$.

We consider 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 that we will make precise in Theorem 1. Along the sequence we consider, the product ϵn tends to a positive constant, so the remainder in (1) is $o_{P_0}(n^{-\frac{1}{2}})$. 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 an expectation under the distribution $\prod_{i=1}^n f_{\beta,\pi}(Y_i)$, 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.* (2016).[1]

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 $\pi_0 \in \Gamma_\epsilon(\gamma_*)$.[2] In the following we focus on locally asymptotically linear estimators that satisfy (1), under the 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.

[1]While in Chernozhukov *et al.* (2016) local robustness is imposed as a substantive restriction on more general moment functions, in our setting (3) and (4) are regularity conditions given unbiasedness.

[2]To see why (1) is a plausible way of imposing asymptotic linearity here, let $\phi(Y_i, \beta_0, \pi_0)$ be the influence

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 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, for example, the OLS estimator under the assumption that $\pi = 0$; e.g., 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 semi-parametric 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

function of $\widehat{\delta}$. Expanding as $n \to \infty$ and $\epsilon \to 0$ we have

$$\widehat{\delta} = \delta_{\beta_0,\pi(\gamma_*)} + \frac{1}{n}\sum_{i=1}^{n} \underbrace{\phi(Y_i, \beta_0, \pi(\gamma_*))}_{=h(Y_i,\beta_0,\gamma_*)} + [\pi_0 - \pi(\gamma_*)]' \underbrace{\left[\nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{\beta_0,\pi(\gamma_*)} \nabla_\pi \phi(Y, \beta_0, \pi(\gamma_*))\right]}_{=0} + o_{P_0}(n^{-\frac{1}{2}} + \epsilon^{\frac{1}{2}}).$$

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 $o_{P_0}(\epsilon^{\frac{1}{2}})$.

estimators to that of our minimum-MSE estimator in our panel data illustration in Section 5.

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.

We assume that $\Gamma_\epsilon(\gamma_*)$ is a convex set. When π is finite-dimensional, for any linear map $u : \Pi \to \mathbb{R}$ we define[3]

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

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 of $d(\pi_0, \pi(\gamma_*))$ 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

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

where

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

for $\|\cdot\|_{\gamma_*}$ the dual norm defined in (6).[4]

Then, the worst-case MSE in $\Gamma_\epsilon(\gamma_*)$ can be expanded as follows, again under appropriate regularity conditions (see Lemma A1 in the appendix),

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta} - \delta_{\beta_0,\pi_0}\right)^2\right] = b_\epsilon(h,\beta_0,\gamma_*)^2 + \frac{\mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h(Y,\beta_0,\gamma_*))}{n} + o(n^{-1}+\epsilon). \tag{9}$$

[3] When π is infinite-dimensional this definition continues to hold for a suitable definition of the scalar product; see Appendix A and Section 3.

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

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

$$h_\epsilon^{\text{MMSE}}(\cdot, \beta_0, \gamma_*) = \underset{h(\cdot,\beta_0,\gamma_*)}{\text{argmin}} \left\{ \epsilon \left\| \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} - \mathbb{E}_{\beta_0,\pi(\gamma_*)} \, h(Y, \beta_0, \gamma_*) \, \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\|_{\gamma_*}^2 \right.$$
$$\left. + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h(Y, \beta_0, \gamma_*))}{n} \right\} \quad \text{subject to (2) and (4).} \tag{10}$$

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

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

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 $\widehat{\gamma}$. 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 $h_\epsilon^{\text{MMSE}}(\cdot, \beta_0, \gamma_*)$ 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 $\widehat{\delta}_\epsilon^{\text{MMSE}}$ 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_ϵ^{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[5]

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

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

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

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

Thus, under the assumption that the parametric reference model is correctly specified, $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is simply the one-step approximation of 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

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

In this limit $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is simply the one-step approximation of 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 $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is a nonlinear interpolation between the one-step MLE approximation of the parametric reference model and the one-step MLE approximation of 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 non point-identified problems, or in problems where the identification of π_0 is irregular. By contrast, $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is always well-defined, since the variance of $h(Y, \beta_0, \gamma_*)$ acts as a sample size-dependent regularization. The form of $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is thus based on both efficiency and robustness. 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 for locally quadratic models in Section 3.

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 Appendix A 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

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

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

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

We establish Theorem 1 in a joint asymptotic where ϵ tends to zero as n tends to infinity and ϵn tends to a finite positive constant. Under our asymptotic 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 explicitly listed in the statement of the theorem. In particular, 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 $\widehat{\gamma}$.[6]

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 pre-specified probability under our local asymptotic. 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

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

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 approaching $1-\alpha$ 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 Appendix A 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, 2016, 2018).[7]

3 Locally quadratic case

In this section we first derive explicit expressions for minimum-MSE estimators in a class of models that have a locally quadratic structure (in Lemmas 1 and 2 below). We then apply

[6]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.

[7]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 $\left|\mathcal{N}\left(1, \widehat{\sigma}_h^2/(n b_\epsilon(h,\widehat{\beta},\widehat{\gamma})^2)\right)\right|$.

these results to parametric and semi-parametric models.

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 $\|u\|^2_{\gamma_*} = u^\top u$, 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 Subsection 3.3), or infinite-dimensional as in semi-parametric models where π is a density (studied in Subsection 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[8]

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

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

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

Lemma 1. *The three following equivalent characterizations of h^{MMSE}_ϵ defined in (10) hold:*

$$\begin{aligned} h^{\text{MMSE}}_\epsilon(y) =& s_{\beta\gamma}(y)' H^{-1}_{\beta\gamma} \nabla_{\beta\gamma}\delta + (\epsilon n)\, \widetilde{s}_\pi(y)^\top \left(\widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h^{\text{MMSE}}_\epsilon(Y) \widetilde{s}_\pi(Y) \right] \right) && (18) \\ =& s_{\beta\gamma}(y)' H^{-1}_{\beta\gamma} \nabla_{\beta\gamma}\delta + (\epsilon n)\, \widetilde{s}_\pi(y)^\top \left(\nabla_\pi \delta - \mathbb{E}\left[h^{\text{MMSE}}_\epsilon(Y) s_\pi(Y) \right] \right) && (19) \\ =& s_{\beta\gamma}(y)' H^{-1}_{\beta\gamma} \nabla_{\beta\gamma}\delta + \widetilde{s}_\pi(y)^\top \left[\widetilde{H}_\pi + (\epsilon n)^{-1} \mathbb{I} \right]^{-1} \widetilde{\nabla}_\pi \delta, && (20) \end{aligned}$$

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

[8] 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 $\overline{\mathcal{T}}$. In Appendix A.1.1 we denote this map by $u^\top = \Omega^{-1}_{\gamma_*} u$.

(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 $h_\epsilon^{\mathrm{MMSE}}$ 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. *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*

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

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 $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ in practice. A second difference comes from the term $\delta(x) - \mathbb{E}_{f_X}\delta(X)$. However this term does not contribute to $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$, since its sample average is zero once we replace f_X by the empirical distribution $\widehat{f}_X$.[9]

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. In Section 4 we will discuss how to choose them in practice.

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

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

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

Corollary 1. *(parametric models)*

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

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

In addition to the "one-step efficient" adjustment $h_0^{\text{MMSE}} = s'_{\beta\gamma}H_{\beta\gamma}^{-1}\nabla_{\beta\gamma}\delta$, the minimum-MSE function h_ϵ^{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. This leads to the following expression, which is identical to (13),

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

where $\widetilde{\mathbb{H}}_\pi^{\dagger}$ denotes the Moore-Penrose (or any other) generalized inverse of $\widetilde{\mathbb{H}}_\pi$. Comparing (22) and Corollary 1 shows that the optimal $\widehat{\delta}_\epsilon^{\text{MMSE}}$ 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

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

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 and Z, and V is normal with zero mean and non-singular 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[10]

$$h_\epsilon^{\text{MMSE}}(y, x, z) = (y - x'\beta_0) x' \Sigma_X^{-1} c \\ - (y - x'\beta_0) \left[(x - Cz) - \Sigma_V \Sigma_X^{-1} x \right]' \left[\left(\Sigma_V - \Sigma_V \Sigma_X^{-1} \Sigma_V \right) + (\epsilon n)^{-1} \sigma^2 I \right]^{-1} \Sigma_V \Sigma_X^{-1} c. \quad (23)$$

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

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

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. When $c'\beta_0$ is identified (that is, 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 essential in order 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 S5 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.

A structural example: conditional cash transfers in Mexico. As an illustration, in Appendix S4 we apply our approach to the structural evaluation of a conditional cash transfer policy in Mexico, the PROGRESA program. This program 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 which were initially randomized

[10] 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 $\mathbb{E}_{f_Z} \widetilde{H}_\pi(Z) = \frac{1}{\sigma^2}(\Sigma_V - \Sigma_V \Sigma_X^{-1} \Sigma_V)$.

out. They compare the predictions of the structural model with the estimated experimental impact. Within a simple static model of education choice, we assess the sensitivity of model-based counterfactual predictions to a particular form of model misspecification under which program participation may have a direct "stigma" effect on the marginal utility of schooling (Wolpin, 2013). We also perform counterfactual predictions in two scenarios – doubling the subsidy amount and implementing an unconditional income transfer – while accounting for the possibility that the reference model is misspecified.

3.4 Semi-parametric models

We now consider semi-parametric models, where the distribution of outcomes Y conditional on unobserved latent variables $A \in \mathcal{A}$ is described parametrically by $Y \,\big|\, A \sim g_{\beta_0}(\cdot \,|\, A)$, 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

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

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 5 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_* are functions $u : \mathcal{A} \mapsto \mathbb{R}$, such as gradients $\nabla_\pi q$ of differentiable functions $q : \Pi \to \mathbb{R}$. For example, the gradient $\nabla_\pi \delta_{\beta,\pi}$ is a cotangent element, which can be represented by the function $\Delta_\beta(\cdot)$.[11] For elements $u, w \in \mathcal{T}$ we define their scalar product by $u^\top w = \mathrm{Cov}_{\pi(\gamma_*)}\,[u(A), w(A)]$; that is, the corresponding squared norm in (6) is $\|u\|^2_{\gamma_*} = \mathrm{Var}_{\pi(\gamma_*)}\,[u(A)]$. One can show that this norm is indeed the dual to a suitable local approximation of 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

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

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

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

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.

Corollary 2. *(semi-parametric models)*

$$\begin{aligned} h_\epsilon^{\mathrm{MMSE}}(y) = & s_{\beta\gamma}(y)' H^{-1}_{\beta\gamma}\, \nabla_{\beta\gamma}\delta + (\epsilon n) \Bigg\{ \mathbb{E}\left[\Delta(A) - \delta - \overline{h}_\epsilon^{\mathrm{MMSE}}(A) \,|\, Y = y\right] \\ & - s_{\beta\gamma}(y)' H^{-1}_{\beta\gamma} \mathbb{E}\left[s_{\beta\gamma}(Y)\left(\Delta(A) - \overline{h}_\epsilon^{\mathrm{MMSE}}(A)\right)\right] \Bigg\}, \end{aligned}$$

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

Covariates. Corollary 2 can readily be generalized to account for conditioning covariates X. We now apply Lemma 2 to provide two generalizations, which we will use in the two

[11] 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 $c = \mathbb{E}_{\pi(\gamma_*)} \Delta_\beta(A)$.

examples in Section 5. In the first one we assume that A and X are independent under π_0. This is the case in our cross-sectional illustration, 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. *(semi-parametric models, independent covariates)*

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

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, 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. *(semi-parametric models, correlated covariates)*

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

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

Remark: connection to the semi-parametric literature. To provide intuition about the form of the solution in the semi-parametric case, 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)

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

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 semi-parametric settings such as panel data, average effects are often only partially identified or not root-n estimable due to ill-posedness.[12] The presence of the Tikhonov penalty $(\epsilon n)^{-1}$ in (25) bypasses these issues by making the operator $[\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}]$ non-singular. By focusing on a shrinking neighborhood of the reference distribution, as opposed to entertaining any possible distribution, our approach avoids issues of non-identification and ill-posedness while guaranteeing MSE-optimality within that neighborhood.[13]

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

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

It follows from Corollary 2 that

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

As ϵ tends to infinity, $\mathbb{W}^{\epsilon}$ approximates the functional differencing projection operator $\mathbb{W} = \mathbb{I}_{\mathcal{Y}} - \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\mathbb{E}^{\dagger}_{\mathcal{A}\,|\,\mathcal{Y}}$, where $\mathbb{E}^{\dagger}_{\mathcal{A}\,|\,\mathcal{Y}}$ 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 semi-parametric efficient estimator of $c'\beta_0$. Yet, the efficient estimator fails to exist when the matrix denominator in (26) is singular.[14] Here the term $(\epsilon n)^{-1}$ acts as a regularization of the functional differencing projection, which makes $h_{\epsilon}^{\text{MMSE}}$ 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 cross-sectional illustration in Section 5. Assuming that β_0 and γ_* are known and $\Delta(A)$ 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

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

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

[12]See, e.g., Chernozhukov *et al.* (2013), Pakes and Porter (2013), Severini and Tripathi (2012), and Bonhomme and Davezies (2017).

[13]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.

[14]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 "non-surjectivity" conditions which may be hard to verify (Bonhomme, 2012).

3.5 Implementation

To implement the method in parametric settings, the researcher needs to compute the score and Hessian of the larger model. In complex economic models this computation will be the main task to implement our approach. 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 semi-parametric models, and computing h_ϵ^{MMSE} based on Corollary 2 requires solving a linear functional system. In Appendix S3 we describe a simulation-based computational method that relies on simple matrix operations. We explain how the same approach can be used to compute confidence intervals. Note that, unlike parametric maximum likelihood estimation or semi-parametric likelihood methods, given initial estimates $\widehat{\beta}$ and $\widehat{\gamma}$ computing minimum-MSE estimators and confidence intervals does not require additional nonlinear optimization.

4 Interpreting the maximal degree of misspecification

In practice, we recommend reporting minimum-MSE estimators and confidence intervals for a range of values of ϵ. Yet, interpretation requires the researcher to assess how large or small a given ϵ-deviation is. This is a fundamental issue in any sensitivity analysis. To facilitate interpretation, here 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. We will illustrate this interpretation through numerical calculations in Section 5.

To proceed, let us focus on the parametric case of Subsection 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 non-central chi-squared distribution with one degree of freedom and non-centrality parameter $\mu = \|\widetilde{H}_\pi^{1/2} v\|\xi$; see, e.g., Van der Vaart (2007, page 237). Here $\widetilde{H}_\pi$ is the usual parametric (projected) Hessian matrix, since Ω is the identity.

Now, for given α and p values, let $\mu(\alpha, p)$ be such that[15]

$$\Pr\left(Z[\mu(\alpha,p)] > \widetilde{c}_\alpha\right) = p.$$

It follows from the previous paragraph that $\mu(\alpha, p) = \|\widetilde{H}_\pi^{1/2} v\|\xi$. Hence, noting that $\mu(\alpha, p)$ is increasing in p and defining

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

we see that taking $\epsilon \geq \epsilon(v)$ ensures that local power in direction v is at least p whenever $\xi/\sqrt{n} \geq \epsilon^{\frac{1}{2}}$. This definition is easy to extend to the general locally-quadratic case of Subsection 3.1.[16]

Setting $\epsilon \geq \epsilon(v)$ is motivated by a desire to calibrate the fear of misspecification of the researcher. When p is large, say 80% or 90%, alternatives in direction v outside the neighborhood $\Gamma_\epsilon(\gamma_*)$ are easy to statistically distinguish from the reference model based on a sample of n observations. Moreover, for fixed α and p the product $\epsilon(v)n$ tends to a constant asymptotically. This aligns well with Huber and Ronchetti (2009, p. 294), who write: "[such] neighborhoods make eminent sense, since the standard goodness-of-fit tests are just able to detect deviations of this order. Larger deviations should be taken care of by diagnostic and modeling, while smaller ones are difficult to detect and should be covered (in the insurance sense) by robustness". A similar logic underlies the calibration strategy developed by Hansen and Sargent (2008).

In order to ensure power larger than p outside the neighborhood in *all* directions v, one could set ϵ to

$$\epsilon_\infty = \sup_{v\,:\,\|v\|=1\,,\,\nabla_{\gamma_*}\pi' v=0} \epsilon(v) = \frac{\mu(\alpha,p)^2}{n\lambda_\infty(\widetilde{H}_\pi)}, \tag{28}$$

where it is sufficient to consider directions that are orthogonal to the directions $\nabla_{\gamma_*}\pi'$ of the reference model, and $\lambda_\infty(\widetilde{H}_\pi)$ denotes the smallest eigenvalue of the matrix or operator $\widetilde{H}_\pi$ projected orthogonally to $\nabla_{\gamma_*}\pi'$. Setting ϵ according to (28) guarantees that all π_0 outside the neighborhood are easy to detect, in agreement with Huber and Ronchetti. However, for ϵ_∞ to be finite, $\widetilde{H}_\pi$ needs to be non-singular, which precludes models with partial or irregular identification.

[15] $\mu(\alpha, p)$ 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.

[16] Specifically, 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 Appendix A.1.1. In the parametric case Ω_{γ_*} is simply the matrix Ω. In the general setup the non-centrality 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 $u \in \mathcal{T}$.

To see this in a concrete example, consider the linear regression model of Subsection 3.3. We obtain, using (28),

$$\epsilon_\infty = \frac{\sigma^2 \mu(\alpha, p)^2}{n\lambda_\infty\left(\Sigma_V - \Sigma_V \Sigma_X^{-1} \Sigma_V\right)}, \tag{29}$$

which is infinite whenever $\Sigma_X - \Sigma_V = C\Sigma_Z C'$ is singular; that is, whenever the IV model is under-identified. 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 more complex examples such as semi-parametric models, the eigenvalues of the infinite-dimensional operator $\widetilde{H}_\pi$ may not be bounded away from zero due to ill-posedness.

As an alternative choice, suppose now that we set ϵ according to

$$\epsilon_1 = \inf_{v\,:\,\|v\|=1} \epsilon(v) = \frac{\mu(\alpha, p)^2}{n\lambda_1(\widetilde{H}_\pi)}, \tag{30}$$

for $\lambda_1(B)$ denotes the maximal eigenvalue of B. ϵ_1 is always finite whenever $\widetilde{H}_\pi \neq 0$, irrespective of the nature of identification. However, setting $\epsilon = \epsilon_1$ only guarantees that parameters outside the neighborhood – those that our robustness adjustments do not insure against – be easily detectable in the most favorable direction. In the IV model, this amounts to ϵ being driven by the most powerful combination of instruments.

In order to perform sensitivity analysis in partially or irregularly identified models, we propose to report several ϵ values given some fixed values of α and p. Specifically, consider 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 k-th largest eigenvalue of B, this logic will lead us to report the first ϵ_k values, where

$$\epsilon_k = \frac{\mu(\alpha, p)^2}{n\lambda_{\mathrm{k}}(\widetilde{H}_\pi)}, \quad \text{for } k = 1, 2, ... \tag{31}$$

In hard-to-estimate models, where point-identification fails or is irregular, there will always exist some directions where our approach will not insure against misspecification. Reporting the "focal" values ϵ_k will then allow the researcher to assess the sensitivity of her results in increasingly large neighborhoods around the reference model. In Appendix S3 we describe how to compute ϵ_k in semi-parametric models using a simulation-based approach.

Finally, it is important to note that here our aim is simply to provide an interpretation for ϵ, and to propose focal values that can be useful to interpret the output of the robustness exercise. An alternative approach, which we do not pursue in this paper, would be to try and develop a data-driven, adaptive choice for ϵ (or a data-driven lower bound on ϵ) under particular assumptions.

Remark: shape of neighborhoods and choice of norm. In addition to ϵ, 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, which we recommend, 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 the same logic as for ϵ, focusing on component-wise 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.[17] In semi-parametric models where Π is a set of densities, we recommend using 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 Illustrations in binary choice models

In this section we apply our approach to two binary choice models, in cross-section and panel data respectively.

5.1 Cross-sectional binary choice

Consider the binary choice model

$$Y = \mathbb{1}\{X'\beta_0 + A \geq 0\}, \tag{32}$$

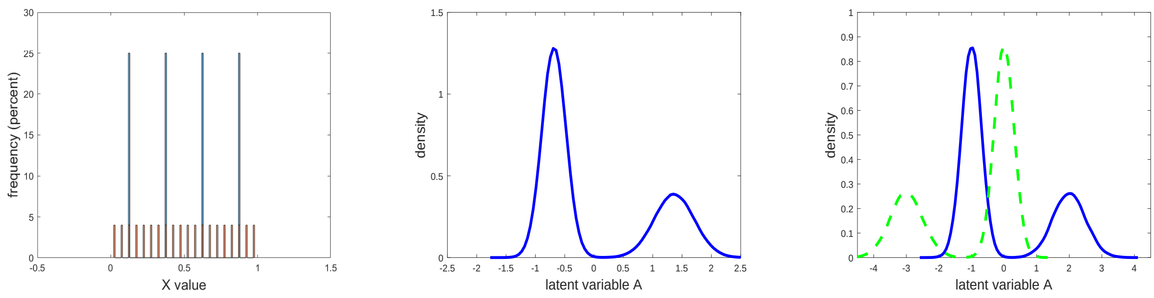
where A follows a distribution π_0, independent of X. We are interested in estimating the prediction function $\delta_{\beta_0,\pi_0} = \mathbb{E}_{\pi_0}[\mathbb{1}\{x_0'\beta_0 + A \geq 0\}]$, at some x_0 not necessarily in the support of X. We focus on the reference specification $A \sim \mathcal{N}(0,1)$, independent of X. We allow for

[17] 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 (2018). While our estimators and confidence intervals remain well-defined in this case, that setting is not locally quadratic.

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 $i = 1, ..., n$.

Figure 1: Distributions of X and A in the binary choice models

(a) Cross-section: X (b) Cross-section: A (c) Panel data: A



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).

The minimum-MSE influence function, in neighborhoods that consist of distributions of A independent of X, 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

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

We compute $h_\epsilon^{\text{MMSE}}(1, X_i)$ and $h_\epsilon^{\text{MMSE}}(0, X_i)$, for $i = 1, ..., n$, based on Corollary 3 by solving a linear system. In this model all conditional expectations are available in closed form, and computation requires no numerical approximation.

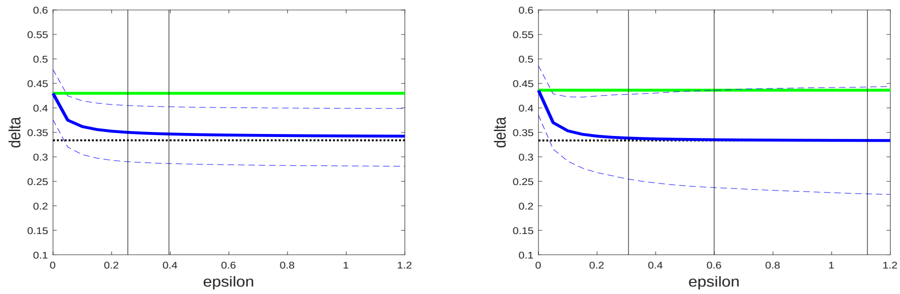
In model (32), under independence between A and X, β_0 and π_0 are point-identified up to scale under sufficiently rich support of X (Manski, 1988). Under such conditions δ_{β_0, π_0} is identified. More generally, it is partially identified. We now set up a simulation where the support of X is discrete, 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 report estimates in data generating processes (DGPs) with a scalar covariate and an intercept, and $\beta_0 = (2, -1)'$. 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

Figure 2: Minimum-MSE estimator in the cross-sectional binary choice model

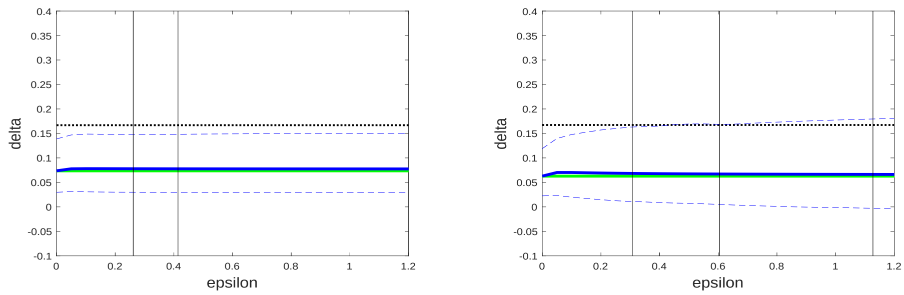
A. Interpolation: $x_0 = (.5, 1)'$

(a) $n_X = 4$ (b) $n_X = 20$



B. Extrapolation: $x_0 = (-.5, 1)'$

(c) $n_X = 4$ (d) $n_X = 20$

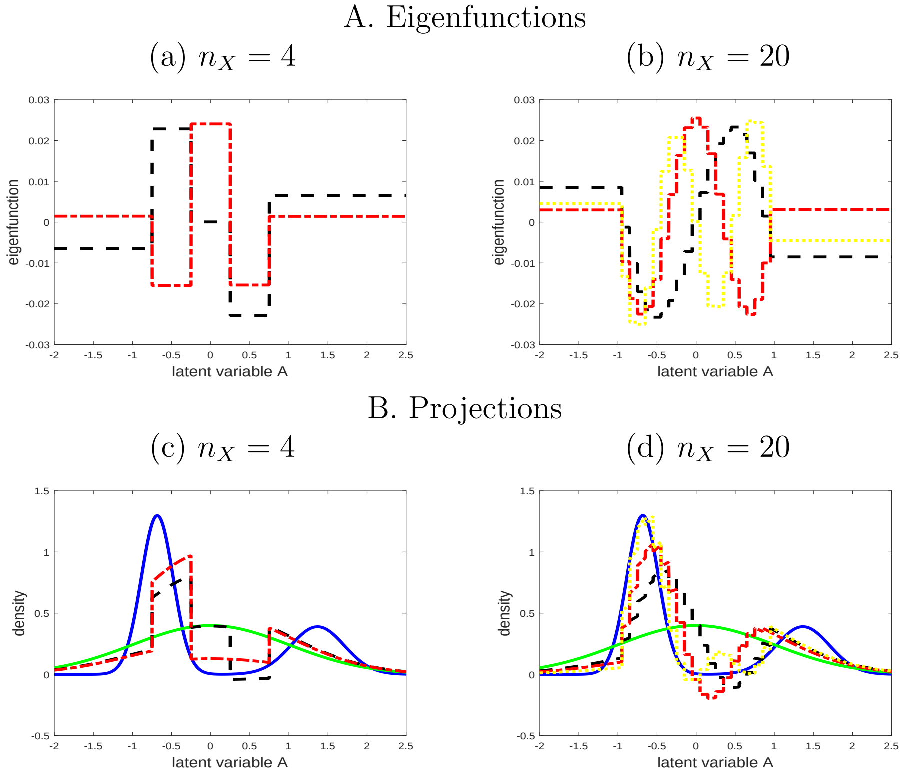


Notes: The solid horizontal line corresponds to the mean probit estimator among 1000 simulations, the solid curve to the mean minimum-MSE estimator (with 2.5% and 97.5% percentiles in dashed), and the dotted horizontal line to the truth. ϵ is reported on the x-axis, and the vertical lines indicate ϵ_k, $k \in \{1, 2, 3\}$. n_X denotes the number of points of support of the first component of X. $n = 500$.

centers 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 1. 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 show the results of the simulation in Figure 2. 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. The vertical lines show our first two focal ϵ values: ϵ_1 and ϵ_2, where we set size and power to $\alpha = 5\%$ and $p = 80\%$. In this setting $\widetilde{H}_\pi$ has only $n_X - 1 = 3$ non-zero eigenvalues (since the $X'\beta$ partition the real line into $n_X + 1$ intervals, and the two

Figure 3: Interpreting ϵ: eigenfunctions of $\widetilde{H}_{\pi}$ in the cross-sectional binary choice model



Notes: In the top panel we report the first 2 (respectively, first 3) non-constant eigenfunctions of $\widetilde{H}_{\pi}$. The first eigenfunction is shown in dashed, the second one in dashed-dotted, and the third one in dotted. In the bottom panel we plot the true and reference densities in solid, as well as the successive approximations using the first, the first two, or the first three eigenfunctions.

elements in β are estimated), two of them corresponding to non-constant eigenfunctions. 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. Note that the degree of misspecification is quite large in the DGP. Indeed, twice the KL divergence between π_0 and π is equal to 1.55, which is larger than ϵ_2. 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. In this case $\widetilde{H}_{\pi}$ has $n_X - 1 = 19$ non-zero eigenvalues. We report the first three focal ϵ values corresponding to non-constant eigenfunctions. We see that the minimum-MSE estimator is virtually unbiased when $\epsilon \geq \epsilon_1$. In 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.

Consider next the lower panel in Figure 2. This "extrapolation" case is very different from the "interpolation" one. Indeed, the data provides 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 $[0, 0.3219]$ (respectively, $[0, 0.2956]$) 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. This suggests that the ability to robustify estimates based on the reference model is limited when one wishes to extrapolate far from the available sample.

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%. 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 all neighborhoods for the range of ϵ that we consider.

Interpreting ϵ. Finally, we use the binary choice model to provide additional intuition about ϵ. Let $\mathcal{U}_k$ denote the span of the first k non-constant eigenfunctions of the operator $\widetilde{H}_\pi$. By construction, any density $\pi_0 \notin \Gamma_{\epsilon_k}(\gamma_*)$ such that $(\pi_0 - \pi(\gamma_*))/\pi(\gamma_*) \in \mathcal{U}_k$ can be "detected" easily, in the sense that the local power of a 5%-likelihood ratio test exceeds 80%.[18] In the upper panel of Figure 3 we plot the eigenfunctions in $\mathcal{U}_k$. Plotting those allows one to visualize the directions along which setting ϵ to either of the ϵ_k's provides power guarantees outside the neighborhood. We see that the eigenfunctions do not vary outside the $[-1, 1]$ interval, where the support of $X'\beta_0$ lies. Within the $[-1, 1]$ interval, the eigenfunctions oscillate and belong to orthogonal bases of functions. To see how well the true π_0 can be approximated using the directions in $\mathcal{U}_k$, in the bottom panel of Figure 3 we report the projection of π_0 onto $\mathcal{U}_k$. We see that outside the $[-1, 1]$ interval the projection is only governed by the reference normal density, reflecting the limited support of X. Within

[18] $\mathcal{U}_k$ consists of cotangent elements that have zero mean under the reference model. Any such $u \in \mathcal{T}$ can be mapped to a direction $v = u \cdot \pi(\gamma_*) \in \overline{\mathcal{T}}$ in the tangent space.

the interval, the approximation to the true bimodal density improves as k increases. At the same time, note that, consistently with our local approach, the approximating functions are not necessarily non-negative.[19]

5.2 Dynamic panel data binary choice

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

$$Y_t = \mathbb{1}\{\beta_0 Y_{t-1} + A + U_t \geq 0\}, \quad t = 1, ..., T, \tag{33}$$

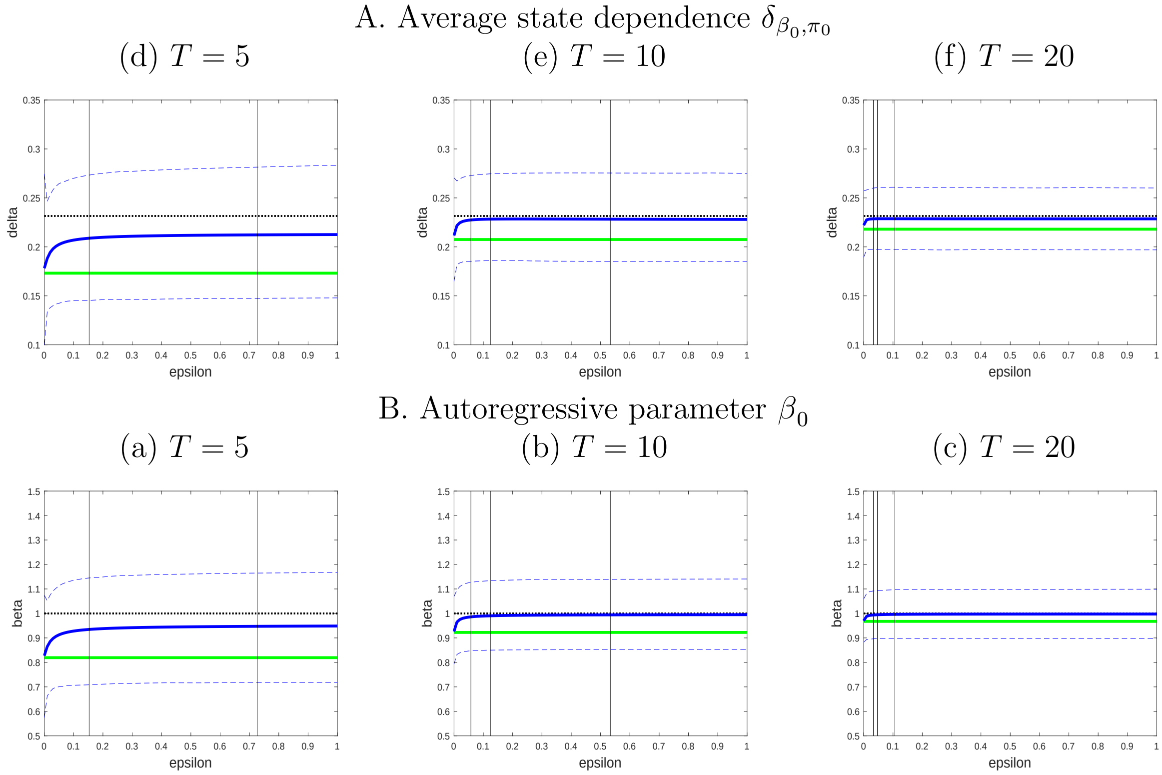
where $U_1, ..., U_T$ are i.i.d. standard normal, independent of A and Y_0. Here Y_0 is observed, so there are effectively $T+1$ time periods. We focus on the average state dependence effect $\delta_{\beta_0,\pi_0} = \mathbb{E}_{\pi_0}[\Phi(\beta_0 + A) - \Phi(A)]$, and we will also report estimates 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 semi-parametrically 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 1. We take $n = 500$, and show the results for $T = 5$, 10, and 20, based on 1000 simulations. The minimum-MSE h function, in neighborhoods that consist of unrestricted joint distributions π_0 of (A, X), 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.

In Figure 4 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

[19] In addition, since we know π_0 in this exercise, we can compute the local power of a 5%-likelihood ratio test in direction $\pi_0 - \pi(\gamma_*)$, for any value of ϵ. We find a power of 0.51 at ϵ_1 and 0.71 at ϵ_2 when X has 4 points of support, and 0.67 at ϵ_1, 0.92 at ϵ_2, and 0.99 at ϵ_3 when X has 20 points of support.

Figure 4: 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. $n = 500$.

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 our focal values ϵ_k. In the top panel of Table 1 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 $\epsilon_1, \epsilon_2, \epsilon_3$.[20] We see the minimum-MSE estimator dominates all other estimators, for all ϵ_k values, when $T = 5$ and $T = 10$. In the bottom panel of Table 1 we show the results for the random-effects MLE and minimum-MSE estimators of β_0. We see similar results as for the case of average state dependence. In this DGP, minimum-MSE estimators achieve bias reduction

[20]The random-effects and empirical Bayes estimators are given by $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\pi(\widehat{\gamma})}[\Phi(\widehat{\beta} + A) - \Phi(A)]$ and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\pi(\widehat{\gamma})}[\Phi(\widehat{\beta} + A) - \Phi(A) \,|\, Y = Y_i]$, respectively. In fixed-lengths panels both estimators are consistent under the parametric reference specification, and the random-effects estimator is efficient. However, the two estimators are generally biased under misspecification, see Bonhomme and Weidner (2019).

Table 1: 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

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

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.

6 Conclusion

We propose a framework for estimation and inference in the presence of model misspecification. Our methods allow 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 can handle parametric and semi-parametric forms of misspecification. It 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. Lastly, 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.

References

[1] Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*.

[2] Andrews, I., M. Gentzkow, and J. M. Shapiro (2018): "On the Informativeness of Descriptive Statistics for Structural Estimates," unpublished manuscript.

[3] Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 132(2), 871–919.

[4] Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

[5] Armstrong, T. B., and M. Kolesár (2016): "Simple and Honest Confidence Intervals in Nonparametric Regression," arXiv preprint arXiv:1606.01200.

[6] Armstrong, T. B., and M. Kolesár (2018): "Sensitivity Analysis Using Approximate Moment Condition Models," arXiv preprint arXiv:1808.07387.

[7] Attanasio, O. P., C. Meghir, and A. 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.

[8] Berger, J., and L. M. Berliner (1986): "Robust Bayes and Empirical Bayes Analysis with ε-Contaminated Priors," *Annals of Statistics*, 461–486.

[9] Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Inference in Semiparametric Models.* Johns Hopkins University Press.

[10] Bonhomme, S. (2012): "Functional Differencing," *Econometrica*, 80(4), 1337–1385.

[11] Bonhomme, S., and L. Davezies (2017): "Panel Data, Inverse Problems, and the Estimation of Policy Parameters," unpublished manuscript.

[12] Bonhomme, S., and Weidner, M. (2019): "Posterior Average Effects," arXiv preprint arXiv:1906.06360.

[13] Bugni, F. A., I. A. Canay, and P. Guggenberger (2012): "Distortions of Asymptotic Confidence Size in Locally Misspecified Moment Inequality Models," *Econometrica*, 80(4), 1741–1768.

[14] Bugni, F. A., and T. Ura (2018): "Inference in Dynamic Discrete Choice Problems under Local Misspecification," to appear in *Quantitative Economics*.

[15] Cattaneo, M. D., R. K. Crump, and M. Jansson (2014): "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives," *Econometric Theory*, 30(1), 176–200.

[16] Chamberlain, G. (1984): "Panel Data", in Griliches, Z. and M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, Elsevier Science, Amsterdam.

[17] Chamberlain, G. (2000): "Econometrics and Decision Theory," *Journal of Econometrics*, 95(2), 255–283.

[18] Chamberlain, G. (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78 (1), 159–168.

[19] Chen, X., E. T. Tamer, and A. Torgovitsky (2011): "Sensitivity Analysis in Semiparametric Likelihood Models," unpublished manuscript.

[20] Cheng, X., and Z. Liao (2015): "Select the Valid and Relevant Moments: An Information-Based LASSO for GMM with Many Moments," *Journal of Econometrics*, 186(2), 443–464.

[21] Chernozhukov, V., J. C. Escanciano, H. Ichimura, and W. K. Newey (2016): "Locally Robust Semiparametric Estimation." arXiv preprint arXiv:1608.00033.

[22] Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535–580.

[23] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, J. (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21(1), C1–C68.

[24] Christensen, T., and B. Connault (2018): "Counterfactual Sensitivity and Robustness," unpublished manuscript.

[25] Claeskens, G., and N. L. Hjort (2003): "The Focused Information Criterion," *Journal of the American Statistical Association*, 98(464), 900–916.

[26] Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly Exogenous," *Review of Economics and Statistics*, 94(1), 260–272.

[27] Dharmadhikari, S. W., and Jogdeo, K. (1969): "Bounds on Moments of Certain Random Variables," *The Annals of Mathematical Statistics*, 40(4), 1506-1509.

[28] Donoho, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 238–270.

[29] Engl, H.W., M. Hanke, and A. Neubauer (2000): *Regularization of Inverse Problems*, Kluwer Academic Publishers.

[30] Fessler, P., and M. Kasy (2018): "How to Use Economic Theory to Improve Estimators," to appear in the *Review of Economics and Statistics.*

[31] Fraser, D. A. S. (1964): "On Local Unbiased Estimation," *Journal of the Royal Statistical Society Series B (Methodological)*, 46–51.

[32] Giacomini, R., and T. Kitagawa (2018): "Robust Bayesian Inference for Set-Identified Models," Cemmap Working Paper No. CWP61/18.

[33] Guggenberger, P. (2012): "On the Asymptotic Size Distortion of Tests when Instruments Locally Violate the Exogeneity Assumption," *Econometric Theory*, 28(2), 387–421.

[34] Gustafson, P. (2000): "Local Robustness in Bayesian Analysis," in *Robust Bayesian Analysis* (pp. 71-88). Springer, New York, NY.

[35] Hahn, J., and J. Hausman (2005): "Estimation with Valid and Invalid Instruments," *Annales d'Economie et de Statistique*, 25–57.

[36] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986): *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics.

[37] Hansen, B. E. (2016): "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 190(1), 115–132.

[38] Hansen, B. E. (2017): "Stein-Like 2SLS Estimator," *Econometric Reviews*, 36(6-9), 840–852.

[39] Hansen, L. P., and M. Marinacci (2016): "Ambiguity Aversion and Model Misspecification: An Economic Perspective," *Statistical Science*, 31(4), 511–515.

[40] Hansen, L. P., and T. J. Sargent (2001): "Robust Control and Model Uncertainty," *American Economic Review*, 91(2), 60–66.

[41] Hansen, L. P., and T. J. Sargent (2008): *Robustness*. Princeton University Press.

[42] Hausman, J., R. Lewis, K. Menzel, and W. Newey (2011): "Properties of the CUE Estimator and a Modification with Moments," *Journal of Econometrics*, 165(1), 45–57.

[43] Heckman, J. 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.

[44] Honoré, B. E., and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–629.

[45] Huber, P. J. (1964): "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35(1), 73–101.

[46] Huber, P. J., and E. M. Ronchetti (2009): *Robust Statistics*. Second Edition. Wiley.

[47] Imbens, G. (2003): "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132.

[48] Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, Infinitesimal Neighborhoods, and Moment Restrictions," *Econometrica*, 81(3), 1185–1201.

[49] Klaassen, C. A. (1987): "Consistent Estimation of the Influence Function of Locally Asymptotically Linear Estimators, *The Annals of Statistics*, 1548–1562.

[50] Leamer, E. (1985): "Sensitivity Analyses Would Help," *American Economic Review*, 75(3), 308–313.

[51] Liao, Z. (2013): "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection," *Econometric Theory*, 29(5), 857–904.

[52] Maasoumi, E. (1978): "A Modified Stein-Like Estimator for the Reduced Form Coefficients of Simultaneous Equations," *Econometrica*, 695–703.

[53] Manski, C. F. (1988): "Identification of Binary Response Models," *Journal of the American statistical Association*, 83.403, 729–738.

[54] Masten, M. A., and A. Poirier (2017): “Inference on Breakdown Frontiers,” arXiv preprint arXiv:1705.04765.

[55] Mueller, U. K. (2012): “Measuring Prior Sensitivity and Prior Informativeness in Large Bayesian Models,” *Journal of Monetary Economics*, 59(6), 581–597.

[56] Mukhin, Y. (2018): “Sensitivity of Regular Estimators.” arXiv preprint arXiv:1805.08883.

[57] Nevo, A., and A. M. Rosen (2012): “Identification with Imperfect Instruments,” *Review of Economics and Statistics*, 94(3), 659–671.

[58] Newey, W. K. (1985): “Generalized Method of Moments Specification Testing,” *Journal of Econometrics*, 29(3), 229–256.

[59] Newey, W. K. (1990): “Semiparametric Efficiency Bounds,” *Journal of Applied Econometrics*, 5(2), 99–135.

[60] Newey, W. K. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 6, 1349–1382.

[61] Neyman, J. (1959): “Optimal Asymptotic Tests of Composite Statistical Hypotheses,” in *Probability and Statistics, the Harald Cramer Volume*, ed. by U. Grenander. Wiley: New York.

[62] Oster, E. (2014): “Unobservable Selection and Coefficient Stability: Theory and Evidence,” to appear in the *Journal of Business & Economic Statistics*.

[63] Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer Verlag Berlin, Heidelberg, New York.

[64] Pakes, A., and J. Porter (2013): “Moment Inequalities for Semiparametric Multinomial Choice with Fixed Effects,” unpublished manuscript.

[65] Rieder, H. (1994): *Robust Asymptotic Statistics*. Springer Verlag, New York, NY.

[66] Rosenbaum, P. R., and D. 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.

[67] Schennach, S. M. (2014): “Entropic Latent Variable Integration via Simulation,” *Econometrica*, 82(1), 345–385.

[68] Todd, P. E., and K. 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.

[69] Todd, P. E., and K. I. Wolpin (2008): “Ex Ante Evaluation of Social Programs,” *Annales d’Economie et de Statistique*, 263–291.

[70] Van der Vaart, A. W. (2007): *Asymptotic Statistics*, Cambridge University Press.

[71] Vidakovic, B. (2000): “Γ-minimax: A Paradigm for Conservative Robust Bayesians,” in *Robust Bayesian analysis* (pp. 241-259). Springer, New York, NY.

[72] Wald, A., 1950: *Statistical Decision Functions*. Wiley, New York.

[73] Watson, J., and C. Holmes (2016): “Approximate Models and Robust Decisions,” *Statistical Science*, 31(4), 465–489.

[74] Wolpin, K. I. (2013): *The limits of Inference Without Theory*. MIT Press.

APPENDIX

A 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 $\pi(\gamma_*)$.[21] 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 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 their scalar product. Notice that in the main text we already introduced cotangent vectors $u \in \mathcal{T}$ and tangent vectors $u^\top \in \overline{\mathcal{T}}$ as their "transpositions", and there we simply wrote $u^\top u$ for their scalar product, which we now write more formally as $\langle u^\top, u\rangle$.

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

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

We assume that there exists a map $\Omega_{\gamma_*} : \overline{\mathcal{T}} \to \mathcal{T}$ such that, for all $v \in \overline{\mathcal{T}}$,

$$\|v\|^2_{\mathrm{ind},\gamma_*} = \left\langle v, \Omega_{\gamma_*} v\right\rangle .$$

We assume that Ω_{γ_*} is invertible, and write $\Omega^{-1}_{\gamma_*} : \mathcal{T} \to \overline{\mathcal{T}}$ for its inverse. The map $\Omega^{-1}_{\gamma_*}$ is exactly the "transposition" map introduced less formally in the main text; that is, for $u \in \mathcal{T}$ we have $u^\top = \Omega^{-1}_{\gamma_*} u \in \overline{\mathcal{T}}$. Thus, our norm on the cotangent space from the main text $\|u\|^2_{\gamma_*} = u^\top u$ can now be written as

$$\|u\|^2_{\gamma_*} = \left\langle \Omega^{-1}_{\gamma_*} u, u \right\rangle .$$

The norm $\|\cdot\|_{\gamma_*}$ is dual to $\|\cdot\|_{\mathrm{ind},\gamma_*}$; that is, we have

$$\|u\|_{\gamma_*} = \sup_{v \in \overline{\mathcal{T}} \setminus \{0\}} \frac{\langle v, u\rangle}{\|v\|_{\mathrm{ind},\gamma_*}}.$$

[21] If Π is a more general manifold (not just an affine space), then the tangent and cotangent spaces depend on the particular value of $\pi \in \Pi$. We then need a connection on the manifold that provides a map between the tangent spaces at $\pi(\gamma_*)$ and $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. All the proofs can be extended to that case, as long as the underlying connection on the manifold is sufficiently smooth. However, this additional formalism is unnecessary to deal with the models discussed in this paper.

Notice also that $\|\cdot\|_{\mathrm{ind},\gamma_*}$, $\|\cdot\|_{\gamma_*}$, Ω_{γ_*}, and $\Omega_{\gamma_*}^{-1}$ could all be defined for general $\pi \in \Pi$, but since we use them only at the reference value $\pi(\gamma_*)$ we index them simply by γ_*.

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

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

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

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

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 of 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 semi-parametric 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

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

where the first inequality is the general relation $H^2(f_{\beta_0,\pi_0}, f_{\beta_0,\pi(\gamma_*)}) \leq D_{\rm KL}(f_{\beta_0,\pi_0}||f_{\beta_0,\pi(\gamma_*)})$ between the squared Hellinger distance H^2 and the Kullback-Leibler divergence $D_{\rm KL}$, 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 semi-parametric case to be twice the Kullback-Leibler divergence $2D_{\rm KL}(\pi_0||\pi(\gamma_*)) = 2\,\mathbb{E}_{\pi_0}\log[\pi_0(A)/\pi(A\,|\,\gamma_*)]$ we find that

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

Thus, the first condition in Assumption A1(iii) is satisfied for those semi-parametric 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 semi-parametric 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 5) 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

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

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 5. The last condition in Assumption A1(iii) could be broken down analogously for semi-parametric 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 $\sup_{y\in\mathcal{Y}}\left\|\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y)\right\|^2_{\gamma_*} = O(1)$.

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 $\widehat{\delta}^{\,\rm MMSE}_{\epsilon} = \delta_{\widehat{\beta},\pi(\widehat{\gamma})} + \frac{1}{n}\sum_{i=1}^n h^{\rm MMSE}_\epsilon(Y_i,\widehat{\beta},\widehat{\gamma})$ 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 $\widehat{\delta}_{\epsilon}^{\text{MMSE}}$ are independent of whether β_0 and γ_* are known or estimated in the construction of $\widehat{\delta}_{\epsilon}^{\text{MMSE}}$ (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) $$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left(\mathbb{E}_{\beta_0,\pi_0} \left\| \begin{pmatrix} \widehat{\beta} \\ \widehat{\gamma} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \gamma_* \end{pmatrix} \right\|^{\chi} \right)^{\frac{1}{\chi}} = O\left(\frac{1}{\sqrt{n}}\right).$$

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

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

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 $\nabla_\eta h_\epsilon^{\text{MMSE}}(y, \beta_0, \gamma_*)$. Since equation (20) in the main text gives an explicit expression for $h_\epsilon^{\text{MMSE}}(y, \beta_0, \gamma_*)$, 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 $h_\epsilon^{\text{MMSE}}(y, \beta_0, \gamma_*)$. 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 $h_\epsilon^{\text{MMSE}}(Y_i, \beta, \gamma)$ 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.

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 A1. *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,*

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0,\pi_0} \right]^2 = b_\epsilon(h_\epsilon, \beta_0, \gamma_*)^2 + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y, \beta_0, \gamma_*))}{n} + o(\epsilon).$$

Lemma A1 provides a formal justification for the worst-case MSE approximation introduced in equation (9) of the main text. In the proof of Theorem 1 we also want to apply Lemma A1 with $h_\epsilon = h_\epsilon^{\rm MMSE}$. Therefore, the following lemma establishes the bounded moment condition on $h_\epsilon^{\rm MMSE}$ required in Lemma A1.

Lemma A2. *Under Assumption A1 the influence function $h_\epsilon^{\rm MMSE}(\cdot, \beta_0, \gamma_*)$ defined in (10) satisfies*

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon^{\rm MMSE}(Y, \beta_0, \gamma_*) \right]^{2+\nu} = O(1).$$

Recall that $\widehat{\delta}_\epsilon^{\rm MMSE} = \widehat{\delta}(h_\epsilon^{\rm MMSE}, \widehat{\beta}, \widehat{\gamma})$. This differs from $\widehat{\delta}(h_\epsilon^{\rm MMSE}, \beta_0, \gamma_*)$, 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 $\widehat{\delta}_\epsilon^{\rm MMSE}$ 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,*

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

Thus, Lemma A3 guarantees that $\widehat{\delta}_\epsilon^{\rm MMSE} = \widehat{\delta}(h_\epsilon^{\rm MMSE}, \beta_0, \gamma_*) + O_{P_0}(1/n)$. 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 $h_\epsilon^{\rm MMSE}$ 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 $\widehat{\delta}_\epsilon^{\rm MMSE} - \widehat{\delta}(h_\epsilon^{\rm MMSE}, \beta_0, \gamma_*)$.

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

$$\begin{aligned} \widehat{\delta}_\epsilon &= \delta_{\beta_0, \pi(\gamma_*)} + \frac{1}{n} \sum_{i=1}^n h_\epsilon(Y_i, \beta_0, \gamma_*) + n^{-1/2} R_n \\ &= \widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) + n^{-1/2} R_n. \end{aligned} \tag{A2}$$

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) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathrm{P}_{\beta_0,\pi_0} \left(|R_n| > \log(n) \right) = o(1),$

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

Then we have

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

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

$$\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)\right] \\
&= \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)-\delta_{\beta_0,\pi_0}+r_\epsilon\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)\right] \\
&= \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)-\delta_{\beta_0,\pi_0}\right)^2 \underbrace{\mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)}_{\leq 1}\right] \\
&\quad + 2\,\mathbb{E}_{\beta_0,\pi_0}\left[r_\epsilon\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)-\delta_{\beta_0,\pi_0}+r_\epsilon\right)\mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)\right] \\
&\quad \underbrace{-\mathbb{E}_{\beta_0,\pi_0}\left[r_\epsilon^2\,\mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)\right]}_{\leq 0} \\
&\leq \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)-\delta_{\beta_0,\pi_0}\right)^2\right] \\
&\quad + 2\,\mathbb{E}_{\beta_0,\pi_0}\left[\underbrace{r_\epsilon\left(\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right)\mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)}_{\leq |r_\epsilon|\, m_n}\right] \\
&\leq \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)-\delta_{\beta_0,\pi_0}\right)^2\right] + 2\, m_n\, \mathbb{E}_{\beta_0,\pi_0}\left|r_\epsilon\right|.
\end{aligned}$$

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

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)\right] \leq \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)-\delta_{\beta_0,\pi_0}\right)^2\right]+o(\epsilon). \tag{A4}$$

By definition h_ϵ^{MMSE} also satisfies the unbiasedness constraint (2). Together with Lemma A2 this implies that h_ϵ^{MMSE} satisfies the conditions on h_ϵ in Lemma A1 with $\kappa = 2+\nu$. Thus, we can apply Lemma A1 with $h_\epsilon = h_\epsilon^{\text{MMSE}}$ to find that

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right)^2\right] = b_\epsilon(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)^2 + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*))}{n} + o(\epsilon). \tag{A5}$$

The function $h_\epsilon^{\text{MMSE}}(\cdot,\beta_0,\gamma_*)$ is defined by the minimization problem (10) in the main text. In other words, $h_\epsilon^{\text{MMSE}}(\cdot,\beta_0,\gamma_*)$ minimizes the objective function $b_\epsilon(h,\beta_0,\gamma_*)^2 + n^{-1}\text{Var}_{\beta_0,\pi(\gamma_*)}(h(Y,\beta_0,\gamma_*))$, 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 $h_\epsilon^{\text{MMSE}}(\cdot,\beta_0,\gamma_*)$ therefore implies that

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

Theorem 1 also imposes all the assumptions on h_ϵ in Lemma A1 . By applying that lemma we thus have

$$b_\epsilon(h_\epsilon,\beta_0,\gamma_*)^2 + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y,\beta_0,\gamma_*))}{n} = \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon,\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right)^2\right] + o(\epsilon). \tag{A7}$$

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

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

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 $\sigma_h^2(\beta_0,\pi_0,\gamma_*) = \text{Var}_{\beta_0,\pi_0} h(Y,\beta_0,\gamma_*)$. *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*

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

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

(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

$$\underbrace{\frac{|\widehat{\delta} - \delta_{\beta_0,\pi_0}| - b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) - \widehat{\sigma}_h\, c_{1-\alpha/2}/\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)/\sqrt{n}}}_{=\text{lhs}}$$
$$\leq \underbrace{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)}\right| - c_{1-\alpha/2} + \widehat{r}_{\beta_0,\pi_0,\gamma_*}}_{=\text{rhs}}, \tag{A9}$$

where

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

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

$$\Pr_{\beta_0,\pi_0}\left[|\widehat{\delta} - \delta_{\beta_0,\pi_0}| \leq b_\epsilon(h, \widehat{\beta}, \widehat{\gamma}) + \frac{\widehat{\sigma}_h}{\sqrt{n}} c_{1-\alpha/2}\right]$$
$$\geq \Pr_{\beta_0,\pi_0}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)}\right| \leq c_{1-\alpha/2} - \widehat{r}_{\beta_0,\pi_0,\gamma_*}\right]. \tag{A10}$$

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

$$\begin{aligned} \left|\widehat{r}_{\beta_0,\pi_0,\gamma_*}\right| \leq \frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \Bigg\{ &\left| \left| \delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} - \mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*) \right| - b_\epsilon(h, \beta_0, \gamma_*) \right| \\ &+ \frac{|\widehat{\sigma}_h - \sigma_h(\beta_0, \pi_0, \gamma_*)|}{\sqrt{n}} c_{1-\alpha/2} + C\,\epsilon^{\frac{1}{2}} \left\| \begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\gamma} - \gamma_* \end{pmatrix} \right\| + |\widehat{R}_{\beta_0,\gamma_*}| \Bigg\}. \end{aligned}$$

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

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

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

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

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. ∎

SUPPLEMENTARY APPENDIX

In Sections S1 and S2 we provide details about the proofs in the paper. In Section S3 we describe our computational approach. In Section S4 we provide an application to the evaluation of a conditional cash transfer program in Mexico. In Section S5 we outline how to extend our approach to models defined by moment restrictions. Lastly, we report additional simulation and estimation results in Section S6.

S1 Complements to main results of Section 2

S1.1 Proof of intermediate lemmas for Theorem 1

The proofs of the Lemmas A1, A2, A3 and A4 are provided in this subsection. Before those proofs it is useful to first establish one additional lemma.

Lemma S1. *Let Assumption A1 hold. Let $q_\epsilon(y)$ and $h_\epsilon(y, \beta_0, \gamma_*)$ be sequences of functions with $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |q(Y)|^\zeta = O(1)$, for some $\zeta > 1$, and $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |h_\epsilon(Y, \beta_0, \gamma_*)|^2 = O(1)$. Then we have*

(i) $$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} \right| = O(\epsilon^{1/2}),$$

(ii) $$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \mathbb{E}_{\beta_0,\pi_0} q_\epsilon(Y) - \mathbb{E}_{\beta_0,\pi(\gamma_*)} q_\epsilon(Y) \right| = O(\epsilon^{1/2}),$$

(iii) $$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \Big| \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) - \left\langle \pi_0 - \pi(\gamma_*), \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\rangle \Big| = o(\epsilon^{1/2}).$$

Proof of Lemma S1. # Part (i): By a mean-value expansion around $\pi(\gamma_*)$ we find

$$\left| \delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} \right| = \left| \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \delta_{\beta_0,\widetilde{\pi}} \right\rangle \right| \le \| \pi_0 - \pi(\gamma_*) \|_{\mathrm{ind},\gamma_*} \left\| \nabla_\pi \delta_{\beta_0,\widetilde{\pi}} \right\|_{\gamma_*},$$

where $\widetilde{\pi}$ is between $\pi(\gamma_*)$ and π_0. Therefore

$$\begin{aligned} \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} \right| &\le \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \| \pi_0 - \pi(\gamma_*) \|_{\mathrm{ind},\gamma_*} \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\| \nabla_\pi \delta_{\beta_0,\pi_0} \right\|_{\gamma_*} \\ &= O(\epsilon^{1/2})\, O(1) = O(\epsilon^{1/2}). \end{aligned}$$

Part (ii): Without loss of generality we assume that $\zeta \le 2$. Let $\xi := \zeta/(\zeta - 1) \ge 2$. We then have

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

where we used that $|a - b| \le |a^c - b^c|^{1/c}$, for any $a, b \ge 0$ and $c \ge 1$, and plugged in $a = f^{1/\xi}_{\beta_0,\pi_0}(y)$, $b = f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)$, and $c = \xi/2$. Thus, the first part of Assumption A1(iii) also implies

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\{ \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y) \right|^\xi dy \right\}^{\frac{1}{\xi}} = O(\epsilon^{1/2}). \tag{S1}$$

Next, we find

$$
\begin{aligned}
&\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left|\mathbb{E}_{\beta_0,\pi_0} q_\epsilon(Y) - \mathbb{E}_{\beta_0,\pi(\gamma_*)} q_\epsilon(Y)\right| \\
&= \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left| \int_{\mathcal{Y}} q_\epsilon(Y) \frac{f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y)}{f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)} \left[f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)\right] dy \right| \\
&\le \left\{ \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} |q_\epsilon(Y)|^{\frac{\xi}{\xi-1}} \left| \frac{f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y)}{f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)} \right|^{\frac{\xi}{\xi-1}} dy \right\}^{\frac{\xi-1}{\xi}} \\
&\qquad\qquad\qquad \times \left\{ \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)\right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\le \xi \left\{ \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} |q_\epsilon(Y)|^{\frac{\xi}{\xi-1}} \left| f_{\beta_0,\pi_0}(y) + f_{\beta_0,\pi(\gamma_*)}(y)\right| dy \right\}^{\frac{\xi-1}{\xi}} \\
&\qquad\qquad\qquad \times \left\{ \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)\right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\le \xi \left\{ 2 \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |q_\epsilon(Y)|^{\zeta} \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)\right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&= o(1),
\end{aligned}
$$

where the first inequality is an application of Hölder's inequality, the second inequality uses that $\left| \frac{f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y)}{f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)} \right|^{\xi/(\xi-1)} \le \xi^{\xi/(\xi-1)} \left[f_{\beta_0,\pi_0}(y) + f_{\beta_0,\pi(\gamma_*)}(y) \right]$,[1] the last line uses that $\kappa = \xi/(\xi-1)$, and the final conclusion follows from our assumptions and (S1).

[1]For $a, b \ge 0$ there exists $c \in [a,b]$ such that by the mean value theorem we have $(a^\xi - b^\xi)/(a-b) = \xi c^{\xi-1} \le \xi \max(a^{\xi-1}, b^{\xi-1})$, and therefore $[(a^\xi - b^\xi)/(a-b)]^{\xi/(\xi-1)} \le \xi^{\xi/(\xi-1)} \max(a^\xi, b^\xi) \le \xi^{\xi/(\xi-1)}(a^\xi + b^\xi)$, which we apply here with $a = f^{1/\xi}_{\beta_0,\pi_0}(y)$ and $b = f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)$.

Part (iii): We have

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y,\beta_0,\gamma_*) \\
&\qquad - \left\langle \pi_0 - \pi(\gamma_*), \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y,\beta_0,\gamma_*) \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\rangle \\
&= \int_{\mathcal{Y}} h_\epsilon(y,\beta_0,\gamma_*) \left[f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y) - \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle f_{\beta_0,\pi(\gamma_*)}(y) \right] dy \\
&= \underbrace{\int_{\mathcal{Y}} h_\epsilon(y,\beta_0,\gamma_*) \left[f^{1/2}_{\beta_0,\pi_0}(y) + f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right] \times \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) - \frac{1}{2} \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right] dy}_{=: a^{(1)}_{\beta_0,\gamma_*,\pi_0}} \\
&\quad + \frac{1}{2} \underbrace{\int_{\mathcal{Y}} h_\epsilon(y,\beta_0,\gamma_*) f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right] dy}_{=: a^{(2)}_{\beta_0,\gamma_*,\pi_0}} .
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality and our assumptions we find that

$$
\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| a^{(1)}_{\beta_0,\gamma_*,\pi_0} \right|^2 \\
&\leq 4 \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} h^2_\epsilon(Y,\beta_0,\gamma_*) \right\} \\
&\qquad \times \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) - \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right\rangle \right]^2 dy \right\} \\
&= O(\epsilon^{1/2}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| a^{(2)}_{\beta_0,\gamma_*,\pi_0} \right|^2 \\
&\leq \left\{ \mathbb{E}_{\beta_0,\pi(\gamma_*)} h^2_\epsilon(Y,\beta_0,\gamma_*) \right\} \\
&\qquad \times \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \|\pi_0 - \pi(\gamma_*)\|^2_{\text{ind},\gamma_*} \int_{\mathcal{Y}} \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|^2_{\gamma_*} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right]^2 dy \right\} \\
&= o(\epsilon).
\end{aligned}
$$

Combining this gives the statement in the lemma. ∎

Proof of Lemma A1. Applying part (ii) of Lemma S1 with $q_\epsilon(y) = h_\epsilon(y,\beta_0,\gamma_*)$ and using the unbiasedness constraint (2) we find that $\mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y,\beta_0,\gamma_*) = o(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Part (i) of Lemma S1 guarantees that $\left|\delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)}\right| = o(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. We therefore have

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} \right]^2 \\
&\quad = \mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y,\beta_0,\gamma_*) \right]^2 - 2 \left(\delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} \right) \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y,\beta_0,\gamma_*) + \left(\delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} \right)^2 \\
&\quad = \mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y,\beta_0,\gamma_*) \right]^2 + o(1),
\end{aligned}
$$

uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Applying part (ii) of Lemma S1 with $q_\epsilon(y) = [h_\epsilon(y, \beta_0, \gamma_*)]^2$ we find that $\mathbb{E}_{\beta_0,\pi_0} [h_\epsilon(Y, \beta_0, \gamma_*)]^2 = \mathbb{E}_{\beta_0,\pi(\gamma_*)} [h_\epsilon(Y, \beta_0, \gamma_*)]^2 + o(1) = \mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y, \beta_0, \gamma_*)) + o(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, where in the last step we have also used that $h_\epsilon(y, \beta_0, \gamma_*)$ satisfies the unbiasedness constraint (2). Therefore,

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y, \beta_0, \gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right]^2 = \mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y, \beta_0, \gamma_*)) + o(1). \tag{S2}$$

Using the unbiasedness constraint again, as well as Lemma S1(iii) and Assumptions A1(ii) and A1(iv) we find

$$\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left|\mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y, \beta_0, \gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right| \\
&= \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left|\left\langle \pi_0 - \pi(\gamma_*), \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) - \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} \right\rangle\right| + o(\epsilon^{1/2}) \\
&= \epsilon^{1/2} \left\|\mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) - \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)}\right\|_{\gamma_*} + o(\epsilon^{1/2}) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} \\
&= b_\epsilon(h_\epsilon, \beta_0, \gamma_*) + o(\epsilon^{1/2}),
\end{aligned} \tag{S3}$$

where in the last step we used the definition of the worst-case bias in (8) of the main text. We furthermore have

$$\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0} \left[\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0,\pi_0}\right]^2 \\
&= \mathbb{E}_{\beta_0,\pi_0} \left(\frac{1}{n} \sum_{i=1}^n h(Y_i, \beta_0, \gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right)^2 \\
&= \left[\mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right]^2 + \frac{1}{n} \mathrm{Var}_{\beta_0,\pi_0} \left[h(Y, \beta_0, \gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right] \\
&= \frac{n-1}{n} \left[\mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*) - \delta_{\beta_0,\pi_0} + \delta_{\beta_0,\pi(\gamma_*)}\right]^2 + \frac{1}{n} \mathbb{E}_{\beta_0,\pi_0} \left[h(Y, \beta_0, \gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right]^2 .
\end{aligned}$$

Taking the supremum of this last result over $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, and then applying (S2) and (S3) gives

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

which is the statement of the lemma. ∎

For the following proof of Lemma A2 it is convenient to introduce some further notation. At the beginning of Section A.1.1 we introduced the vector norms $\|\cdot\|_{\mathrm{ind},\gamma_*}$, $\|\cdot\|_{\gamma_*}$ and $\|.\|$ on $\overline{\mathcal{T}}$, $\mathcal{T}$ and $\mathbb{R}^{\dim\beta+\dim\gamma}$. Those vector norms induce natural norms on any maps between $\overline{\mathcal{T}}$, $\mathcal{T}$ and $\mathbb{R}^{\dim\beta+\dim\gamma}$. With a slight abuse of notation we denote all those norms simply by $\|.\|_{\gamma_*}$. In particular, for $\Omega_{\gamma_*}^{-1} : \mathcal{T} \to \overline{\mathcal{T}}$ we have

$$\left\|\Omega_{\gamma_*}^{-1}\right\|_{\gamma_*} := \sup_{u \in \mathcal{T}\setminus\{0\}} \frac{\|\Omega_{\gamma_*}^{-1} u\|_{\mathrm{ind},\gamma_*}}{\|u\|_{\gamma_*}} = \sup_{u \in \mathcal{T}\setminus\{0\}} \frac{\left\langle \Omega_{\gamma_*}^{-1} u, , u \right\rangle^{1/2}}{\|u\|_{\gamma_*}} = 1, \tag{S4}$$

and for $H_{\pi,\beta\gamma} : \mathbb{R}^{\dim\beta+\dim\gamma} \to \mathcal{T}$ defined in Section 3.1 we have

$$\left\| H_{\pi,\beta\gamma} \right\|_{\gamma_*} := \sup_{w\in\mathbb{R}^{\dim\beta+\dim\gamma}\setminus\{0\}} \frac{\|H_{\pi,\beta\gamma}w\|_{\gamma_*}}{\|w\|} = \sup_{v\in\overline{\mathcal{T}}\setminus\{0\}} \sup_{w\in\mathbb{R}^{\dim\beta+\dim\gamma}\setminus\{0\}} \frac{\langle v, H_{\pi,\beta\gamma}w\rangle}{\|v\|_{\text{ind},\gamma_*}\,\|w\|}.$$

Using Assumption A1(v) and the Cauchy-Schwarz inequality we find that

$$\begin{aligned}\left\| H_{\pi,\beta\gamma} \right\|_{\gamma_*} &= \left\| \mathbb{E}_{\beta_0,\pi(\gamma_*)} \left\{ \left[\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y)\right] \left[\nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(Y)\right]' \right\} \right\|_{\gamma_*} \\ &\le \left[\mathbb{E}_{\beta_0,\pi(\gamma_*)} \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\|^2_{\gamma_*} \right]^{1/2} \left[\mathbb{E}_{\beta_0,\pi(\gamma_*)} \left\| \nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\|^2 \right]^{1/2} \\ &= O(1). \end{aligned} \tag{S5}$$

Proof of Lemma A2. Equation (20) in Lemma 1 in the main text provides an explicit solution for $h^{\text{MMSE}}_\epsilon(y,\beta_0,\gamma_*)$, which in the notation of this appendix can be written as

$$\begin{aligned} h^{\text{MMSE}}_\epsilon(y,\beta_0,\gamma_*) &= \left[\nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)}\right]' H^{-1}_{\beta\gamma} \left[\nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(y)\right] \\ &\quad + \left\langle \left[\widetilde{H}_\pi\, \Omega_{\gamma_*} + (\epsilon n)^{-1} \Omega_{\gamma_*} \right]^{-1} \widetilde{\nabla}_\pi \delta_{\beta_0,\pi(\gamma_*)} \,,\, \widetilde{\nabla}_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle, \end{aligned}$$

where $\widetilde{\nabla}_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) = \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) - H_{\pi,\beta\gamma} H^{-1}_{\beta\gamma}\, \nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(y)$ and $\widetilde{\nabla}_\pi \delta_{\beta_0,\pi(\gamma_*)} = \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} - H_{\pi,\beta\gamma} H^{-1}_{\beta\gamma}\, \nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)}$. We thus have

$$\begin{aligned} \left| h^{\text{MMSE}}_\epsilon(y,\beta_0,\gamma_*) \right| &\le \left\| \nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)} \right\| \left\| H^{-1}_{\beta\gamma} \right\| \left\| \left[\nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(y)\right] \right\| \\ &\quad + (\epsilon n) \left\| \Omega^{-1}_{\gamma_*} \right\|_{\gamma_*} \left\| \widetilde{\nabla}_\pi \delta_{\beta_0,\pi(\gamma_*)} \right\|_{\gamma_*} \left\| \widetilde{\nabla}_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|_{\gamma_*}, \end{aligned}$$

where we used that $\left\| \left[\widetilde{H}_\pi\, \Omega_{\gamma_*} + (\epsilon n)^{-1} \Omega_{\gamma_*} \right]^{-1} \right\|_{\gamma_*} \le (\epsilon n) \left\| \Omega^{-1}_{\gamma_*} \right\|_{\gamma_*}$, because both $\widetilde{H}_\pi\, \Omega_{\gamma_*}$ and Ω_{γ_*} are positive semi-definite. We furthermore have

$$\begin{aligned} \left\| \widetilde{\nabla}_\pi \delta_{\beta_0,\pi(\gamma_*)} \right\|_{\gamma_*} &\le \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|_{\gamma_*} + \left\| H_{\pi,\beta\gamma} \right\|_{\gamma_*} \left\| H^{-1}_{\beta\gamma} \right\| \left\| \nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|, \\ \left\| \widetilde{\nabla}_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|_{\gamma_*} &\le \left\| \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} \right\|_{\gamma_*} + \left\| H_{\pi,\beta\gamma} \right\|_{\gamma_*} \left\| H^{-1}_{\beta\gamma} \right\| \left\| \nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)} \right\|. \end{aligned}$$

Combining those inequalities with our Assumption A1(ii) and (v) as well as the results (S4) and (S5) above we find that

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[h^{\text{MMSE}}_\epsilon(Y,\beta_0,\gamma_*) \right]^{2+\nu} = O(1).$$

∎

Proof of Lemma A3. Let $\eta = (\beta',\gamma')'$, $\widehat{\eta} := (\widehat{\beta}', \widehat{\gamma}')'$, and $\eta_* := (\beta_0', \gamma_*')'$. By a Taylor expansion in η around η_* we find that

$$\begin{aligned}
\widehat{\delta}_\epsilon^{\text{MMSE}} &= \delta_{\widehat{\beta},\pi(\widehat{\gamma})} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\text{MMSE}}(Y_i,\widehat{\beta},\widehat{\gamma}) \\
&= \delta_{\beta_0,\pi(\gamma_*)} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\text{MMSE}}(Y_i,\beta_0,\gamma_*) \\
&\quad \underbrace{(\widehat{\eta}-\eta_*)'\left[\nabla_\eta \delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*)\right]}_{=r^{(1)}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta_*)'\frac{1}{n}\sum_{i=1}^n \left[\nabla_\eta h_\epsilon^{\text{MMSE}}(Y_i,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y_i,\beta_0,\gamma_*)\right]}_{=r^{(2)}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta_*)'\left[\mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*)\right]}_{=r^{(3)}} \\
&\quad + \frac{1}{2}\underbrace{(\widehat{\eta}-\eta_*)'\left[\frac{1}{n}\sum_{i=1}^n \nabla^2_{\eta\eta'} h_\epsilon^{\text{MMSE}}(Y_i,\widetilde{\beta},\widetilde{\gamma})\right](\widehat{\eta}-\eta_*)}_{=r^{(4)}},
\end{aligned} \tag{S6}$$

where $\widetilde{\eta} = (\widetilde{\beta}',\widetilde{\gamma}')'$ is a value between $\widehat{\eta}$ and η_*. Our constraints (2) and (4) guarantee that $\nabla_\eta \delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*) = 0$; that is, we have $r^{(1)} = 0$. Using Assumption A2 and the Cauchy-Schwarz inequality we furthermore find

$$\begin{aligned}
&\left(\mathbb{E}_{\beta_0,\pi_0}\left|r^{(2)}\right|\right)^2 \\
&\leq \mathbb{E}_{\beta_0,\pi_0}\|\widehat{\eta}-\eta_*\|^2 \;\mathbb{E}_{\beta_0,\pi_0}\left\|\frac{1}{n}\sum_{i=1}^n\left[\nabla_\eta h_\epsilon^{\text{MMSE}}(Y_i,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y_i,\beta_0,\gamma_*)\right]\right\|^2 \\
&\leq \mathbb{E}_{\beta_0,\pi_0}\|\widehat{\eta}-\eta_*\|^2 \;\frac{1}{n}\mathbb{E}_{\beta_0,\pi_0}\left\|\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*)\right\|^2 = O\left(\frac{1}{n^2}\right),
\end{aligned}$$

uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, where in the second step we have used the independence of Y_i across i. Similarly, we have

$$\begin{aligned}
\left(\mathbb{E}_{\beta_0,\pi_0}\left|r^{(3)}\right|\right)^2 &\leq \mathbb{E}_{\beta_0,\pi_0}\|\widehat{\eta}-\eta_*\|^2\left\|\mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*)\right\|^2 \\
&= O\left(\frac{1}{n}\right)O(\epsilon) = O\left(\frac{1}{n^2}\right),
\end{aligned}$$

uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, where we have used that

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left\|\mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*)\right\| = O(\epsilon^{1/2}),$$

which follows from Assumptions A1(iii) and A2(ii) by using the proof strategy of part (ii) of Lemma S1. Finally, applying Hölder's inequality we have

$$\begin{aligned}
\mathbb{E}_{\beta_0,\pi_0}\left|r^{(4)}\right| &\leq \mathbb{E}_{\beta_0,\pi_0}\left[\|\widehat{\eta}-\eta_*\|^2\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\eta\eta'}h_\epsilon^{\text{MMSE}}(Y_i,\widetilde{\beta},\widetilde{\gamma})\right\|\right] \\
&\leq \left\{\mathbb{E}_{\beta_0,\pi_0}\|\widehat{\eta}-\eta_*\|^\chi\right\}^{\frac{2}{\chi}}\left\{\mathbb{E}_{\beta_0,\pi_0}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\eta\eta'}h_\epsilon^{\text{MMSE}}(Y_i,\widetilde{\beta},\widetilde{\gamma})\right\|^{\frac{\chi}{\chi-2}}\right\}^{\frac{\chi-2}{\chi}} \\
&= O\left(\frac{1}{n}\right),
\end{aligned}$$

uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, where we have used Assumption A2(iii). We have thus shown that

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\left|r^{(1)}+r^{(2)}+r^{(3)}+\frac{1}{2}r^{(4)}\right| = O\left(\frac{1}{n}\right),$$

which together with (S6) gives the statement of the lemma. ∎

The proof of the next lemma uses the following theorem of Petrov (1975), which generalizes the Berry-Esseen theorem to sample averages of random variables without a third moment.

Theorem S11 (Theorem 5 on p. 112 in Petrov 1975)**.** *Let $X_1,\ldots,X_n$ be independent random variables, such that $\mathbb{E}X_j = 0$, $\mathbb{E}(X_j^2 g(|X_j|)) < \infty$ for $j = 1,\ldots,n$, and for some function $g : [0,\infty) \to [0,\infty)$ such that both $g(x)$ and $x/g(x)$ are non-decreasing for $x > 0$. We write*

$$\sigma_j^2 = \mathbb{E}X_j^2, \qquad B_n = \sum_{j=1}^{n}\sigma_j^2, \qquad F_n(x) = \Pr\left(B_n^{-1/2}\sum_{j=1}^{n}X_j < x\right).$$

Then there exists an absolute constant $A > 0$ such that

$$\sup_x |F_n(x)-\Phi(x)| \leq \frac{A}{B_n g(\sqrt{B_n})}\sum_{j=1}^{n}\mathbb{E}\left(X_j^2 g(X_j)\right).$$

Proof of Lemma A4. # Preliminaries: We first establish some preliminary results on the sample averages of

$$\widetilde{h}_\epsilon(Y_i,\beta_0,\gamma_*,\pi_0) := h_\epsilon(Y_i,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi_0}h_\epsilon(Y_i,\beta_0,\gamma_*).$$

According to our assumptions the $\widetilde{h}_\epsilon(Y_i,\beta_0,\gamma_*,\pi_0)$ are independent random variables with zero mean and finite absolute moments of order $\kappa > 2$, under $P_0 = P(\beta_0,\pi_0)$. By applying the result in Dharmadhikari and Jogdeo (1969) we thus find that[2]

$$\mathbb{E}_{\beta_0,\pi_0}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{h}_\epsilon(Y_i,\beta_0,\gamma_*,\pi_0)\right|^\kappa \leq C_\kappa\,\mathbb{E}_{\beta_0,\pi_0}\left|\widetilde{h}_\epsilon(Y_i,\beta_0,\gamma_*,\pi_0)\right|^\kappa,$$

[2] This result is an extension of the Bahr-Esseen inequality to moments larger than two. See also inequality number 16 on p. 60 of Petrov (1975).

where the constant $C_\kappa > 0$ only depends on κ. Through a combination of the Minkowski and Hölder's inequalities we find that our assumption $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |h_\epsilon(Y, \beta_0, \gamma_*)|^\kappa = O(1)$ also guarantees $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left|\widetilde{h}_\epsilon(Y, \beta_0, \gamma_*)\right|^\kappa = O(1)$. We therefore obtain that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left(\mathbb{E}_{\beta_0,\pi_0} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0) \right|^\kappa \right)^{\frac{1}{\kappa}} = O(1). \tag{S7}$$

Next, we apply Theorem 5 of Chapter V in Petrov (1975), which is restated above as Theorem S11, with X_i equal to $\widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0)$ and $g(x) = x^{\min\{1,\kappa-2\}}$ to find that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \sup_{x \in \mathbb{R}} \left| \mathrm{P}_{\beta_0,\pi_0} \left(\frac{\sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0)}{\sqrt{n}\, \sigma(\beta_0, \gamma_*, \pi_0)} \leq x \right) - \Phi(x) \right| = o(1),$$

where $\sigma^2(\beta_0, \gamma_*, \pi_0) = \mathbb{E}_{\beta_0,\pi_0} \widetilde{h}_\epsilon^2(Y_i, \beta_0, \gamma_*, \pi_0)$. This, in particular, implies that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathrm{P}_{\beta_0,\pi_0} \left(\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0) \right| > \log(n) \right) = o(1). \tag{S8}$$

By an application of Hölder's inequality we find that (S7) and (S8) also imply

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*) \right)^2 \mathbb{1} \left(\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*) \right| > \log n \right) \right] = o(1). \tag{S9}$$

Finally, we notice that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} + \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y, \beta_0, \gamma_*) \right| = O(\epsilon^{1/2}), \tag{S10}$$

which follows by applying part (i) and (ii) of Lemma S1 with $q_\epsilon(y) = h_\epsilon(y, \beta_0, \gamma_*)$ and noting that $\mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) = 0$ by the unbiasedness constraint (2).

\# Main result of the Lemma A4: Having established those preliminary results, we now derive the statement of the lemma. Define

$$\begin{aligned} k_n &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n h_\epsilon(Y_i, \beta_0, \gamma_*) + \sqrt{n} \left[\delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} \right] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*) + \sqrt{n} \left[\delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} + \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y_i, \beta_0, \gamma_*) \right]. \end{aligned}$$

The decomposition of $\widehat{\delta}_\epsilon$ in (A2) can then be rewritten as

$$\sqrt{n} \left(\widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0} \right) = k_n + R_n.$$

We have

$$
\begin{aligned}
n\,\mathbb{E}_{\beta_0,\pi_0}&\left[\left(\widehat{\delta}_\epsilon-\delta_{\beta_0,\pi_0}\right)^2\mathbb{1}\left(\left|\widehat{\delta}_\epsilon-\delta_{\beta_0,\pi_0}\right|\leq m_n\right)\right]\\
&=\mathbb{E}_{\beta_0,\pi_0}\left[(k_n+R_n)^2\,\mathbb{1}\left(|k_n+R_n|\leq n^{1/2}\,m_n\right)\right]\\
&=\mathbb{E}_{\beta_0,\pi_0}k_n^2-\underbrace{\mathbb{E}_{\beta_0,\pi_0}\left[k_n^2\,\mathbb{1}\left(|k_n+R_n|>n^{1/2}\,m_n\right)\right]}_{=\text{term I}}\\
&\qquad\qquad+\underbrace{\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2+2k_n\,R_n)\mathbb{1}\left(|k_n+R_n|\leq n^{1/2}\,m_n\right)\right]}_{=\text{term II}}.
\end{aligned}
$$

Thus, Lemma A4 is proved if we can show that term I is $o(1)$, and that term II is larger or equal to minus $o(1)$, both uniformly over $\pi_0\in\Gamma_\epsilon(\gamma_*)$. For term I we use Hölder's inequality to obtain that

$$
\begin{aligned}
&\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\left[k_n^2\,\mathbb{1}\left(|k_n+R_n|>n^{1/2}\,m_n\right)\right]\\
&\leq\left\{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left(\mathbb{E}_{\beta_0,\pi_0}|k_n|^\kappa\right)^{\frac{2}{\kappa}}\right\}\left\{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left[\mathbb{E}_{\beta_0,\pi_0}\mathbb{1}\left(|k_n+R_n|>n^{1/2}\,m_n\right)\right]^{\frac{\kappa-2}{\kappa}}\right\}\\
&\leq\left\{\underbrace{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left(\mathbb{E}_{\beta_0,\pi_0}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\widetilde{h}_\epsilon(Y_i,\beta_0,\gamma_*)\right|^\kappa\right)^{\frac{2}{\kappa}}}_{=O(1)}\right.\\
&\qquad\qquad\left.+\underbrace{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left(n^{1/2}\left|\delta_{\beta_0,\pi(\gamma_*)}-\delta_{\beta_0,\pi_0}+\mathbb{E}_{\beta_0,\pi_0}h_\epsilon(Y_i,\beta_0,\gamma_*)\right|\right)}_{=O(1)}\right\}\\
&\quad\times\left\{\left[\underbrace{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\mathbb{1}\left(|k_n|>\frac{1}{2}\,n^{1/2}\,m_n\right)}_{=o(1)}+\underbrace{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\mathbb{1}\left(|R_n|>\frac{1}{2}\,n^{1/2}\,m_n\right)}_{=o(1)}\right]^{\frac{\kappa-2}{\kappa}}\right\}\\
&=o(1),
\end{aligned}
$$

where we also used the definition of k_n together with the triangle inequality, and we employed (S7), (S8) and (S10) and Assumption (ii) of the lemma, together with our assumption that $n^{1/2}\,m_n\gg\log(n)$ as $n\to\infty$.

Next, for term II we use that $R_n^2+2k_n\,R_n$ is positive whenever $|R_n|>2\,|k_n|$ to obtain that

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2+2k_n\,R_n)\mathbb{1}\left(|k_n+R_n|\le n^{1/2}\,m_n\right)\right]\\
&\quad=\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2+2k_n\,R_n)\mathbb{1}\left(|k_n+R_n|\le n^{1/2}\,m_n\right)\mathbb{1}\left(|R_n|\le 2\,|k_n|\right)\right]\\
&\qquad+\underbrace{\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2+2k_n\,R_n)\mathbb{1}\left(|k_n+R_n|\le n^{1/2}\,m_n\right)\mathbb{1}\left(|R_n|> 2\,|k_n|\right)\right]}_{\ge 0}\\
&\quad\ge\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2+2k_n\,R_n)\mathbb{1}\left(|k_n+R_n|\le n^{1/2}\,m_n\right)\mathbb{1}\left(|R_n|\le 2\,|k_n|\right)\right]\\
&\quad\ge -2\,\mathbb{E}_{\beta_0,\pi_0}\Big[|k_n|\,|R_n|\;\mathbb{1}\left(|R_n|\le 2\,|k_n|\right)\Big]\\
&\quad\ge -2\,\left\{\mathbb{E}_{\beta_0,\pi_0}\,k_n^2\right\}^{1/2}\left\{\mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n|\le 2\,|k_n|\right)\right]\right\}^{1/2}
\end{aligned}
$$

where in the last step we also used the Cauchy-Schwarz inequality. Our preliminary results (S7) and (S10) imply that $\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\,k_n^2=O(1)$. Furthermore we have

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n|\le 2\,|k_n|\right)\right]\\
&\quad=\mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n|\le 2\,|k_n|\right)\mathbb{1}\left(|k_n|\le \log n\right)\right]\\
&\qquad+\mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n|\le 2\,|k_n|\right)\mathbb{1}\left(|k_n|> \log n\right)\right]\\
&\quad\le\mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n|\le 2\,\log n\right)\right]+4\,\mathbb{E}_{\beta_0,\pi_0}\left[k_n^2\,\mathbb{1}\left(|k_n|> \log n\right)\right]\\
&\quad=o(1),
\end{aligned}
$$

uniformly over $\pi_0\in\Gamma_\epsilon(\gamma_*)$, where we used (S9) and Assumption (v) of the lemma. We thus conclude that term II indeed satisfies

$$
\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left\{-\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2+2k_n\,R_n)\mathbb{1}\left(|k_n+R_n|\le n^{1/2}\,m_n\right)\right]\right\}\le o(1).
$$

Combining the above gives the statement of the lemma. ∎

S1.2 Lemma 1

Before deriving the equivalent characterizations of $h_\epsilon^{\rm MMSE}(y,\beta_0,\gamma_*)$ given in the lemma we note that the optimization problem (10) that defines $h_\epsilon^{\rm MMSE}(y,\beta_0,\gamma_*)$ has a unique solution (up to possible deviations on a measure zero set of y's, which are irrelevant for our purposes). This uniqueness follows, because under the unbiasedness constraint (2) we have $\mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h(Y,\beta_0,\gamma_*))=\mathbb{E}_{\beta_0,\pi(\gamma_*)}h^2(Y,\beta_0,\gamma_*)$, which is quadratic and strictly convex in $h(y,\beta_0,\gamma_*)$, while all other components of the objective function and constraints in (10) are linear in $h(y,\beta_0,\gamma_*)$.

Equation (18). Using simplified notation here, our goal is to find the function $h(y)=h(y,\beta_0,\gamma_*)$ that minimizes

$$
\mathbb{E}h^2(Y)+(\epsilon n)\,\left\{\nabla_\pi\delta-\mathbb{E}\left[h(Y)s_\pi(Y)\right]\right\}^\top\left\{\nabla_\pi\delta-\mathbb{E}\left[h(Y)s_\pi(Y)\right]\right\},
$$

subject to the constraints $\mathbb{E}h(Y)=0$ and $\mathbb{E}\,h(Y)s_{\beta\gamma}(Y)=\nabla_{\beta\gamma}\delta$.

Using the latter constraint and the definition of $\widetilde{\nabla}_\pi$ we can equivalently rewrite the objective function as

$$\mathbb{E}h^2(Y) + (\epsilon n) \left\{ \widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h(Y)\widetilde{s}_\pi(Y)\right] \right\}^\top \left\{ \widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h(Y)\widetilde{s}_\pi(Y)\right] \right\} \\ + 2\left\{ \nabla_{\beta\gamma}\delta - \mathbb{E}\left[h(Y)s_{\beta\gamma}(Y)\right] \right\}' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma}\delta.$$

The unconstrained minimizer of this rewritten quadratic objective function satisfies the first-order condition

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

and because $\mathbb{E}s_{\beta\gamma}(Y) = 0$, $\mathbb{E}\widetilde{s}_\pi(Y) = 0$, and $\mathbb{E}[s_{\beta\gamma}(Y)s_{\beta\gamma}(Y)'] = H_{\beta\gamma}$, we find that this unconstrained minimizer already satisfies both constraints $\mathbb{E}h(Y) = 0$ and $\mathbb{E}h(Y)s_{\beta\gamma}(Y) = \nabla_{\beta\gamma}\delta$, and is therefore also the constrained minimizer that we wanted to derive.

Equation (19). Note that by (18) we have $h_\epsilon^{\text{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma}\delta + \widetilde{s}_\pi(y)^\top u$, for some $u \in \mathcal{T}$, and one can easily verify that this implies that $\widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y)\widetilde{s}_\pi(Y)\right]$ is equal to the same expression with $\widetilde{s}_\pi$ replaced by s_π.

Equation (20). We have already shown that equation (18) is the FOC of the minimization problem (10). We now want to show that the solution for $h_\epsilon^{\text{MMSE}}(y)$ given in equation (20) satisfies the FOC (18), which implies that it solves (10). Equation (18) can be rewritten as

$$h_\epsilon^{\text{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma}\delta + (\epsilon n)\, \widetilde{s}_\pi(y)^\top u, \quad u := \widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y)\widetilde{s}_\pi(Y)\right]. \tag{S11}$$

Plugging the expression for $h_\epsilon^{\text{MMSE}}(y)$ given by equation (20) into this definition of u and using that $\mathbb{E}\left[\widetilde{s}_\pi(Y)\widetilde{s}_\pi(Y)^\top\right] = \widetilde{H}_\pi$, and $\mathbb{E}\left[\widetilde{s}_\pi(Y)s_{\beta\gamma}(Y)'\right] = 0$, we find that (20) implies that

$$\begin{aligned} u &= \widetilde{\nabla}_\pi \delta - \widetilde{H}_\pi \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I} \right]^{-1} \widetilde{\nabla}_\pi \delta \\ &= \left\{ \mathbb{I} - \widetilde{H}_\pi \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I} \right]^{-1} \right\} \widetilde{\nabla}_\pi \delta \\ &= \left\{ \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I} \right] \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I} \right]^{-1} - \widetilde{H}_\pi \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I} \right]^{-1} \right\} \widetilde{\nabla}_\pi \delta \\ &= (\epsilon n)^{-1} \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I} \right]^{-1} \widetilde{\nabla}_\pi \delta. \end{aligned}$$

This expression for u makes the first equation in (S11) equivalent to (20). Therefore, we have shown that $h_\epsilon^{\text{MMSE}}(y)$ as given by (20) indeed solves (18), and therefore also our optimization problem in (10).

S1.3 Lemma 2

Our goal is to choose the function $h(\cdot,\cdot,\beta,\gamma,f_X)$ such that the worst-case mean squared error

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

is minimized for small values of ϵ, subject to unbiasedness under the reference model, and also subject to local robustness constraints to account for the fact that β_0, γ_* and f_X are estimated from the sample.

Unbiasedness is

$$\mathbb{E}_{f_X}\,\mathbb{E}_{\beta_0,\pi(\gamma_*)}\,h(Y,X,\beta_0,\gamma_*,f_X)=0, \tag{S12}$$

while local robustness is

$$\begin{aligned}&\mathbb{E}_{f_X}\,\mathbb{E}_{\beta_0,\pi(\gamma_*)}\,h(Y,X,\beta_0,\gamma_*,f_X)\,\nabla_{\beta\gamma}\log f_{\beta_0,\pi(\gamma_*)}(Y\,|\,X)=\mathbb{E}_{f_X}\,\nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)}(X),\\&\mathbb{E}_{\beta_0,\pi(\gamma_*)}\,\left[h(Y,X,\beta_0,\gamma_*,f_X)\,|\,X=x\right]=\delta_{\beta_0,\pi(\gamma_*)}(x)-\mathbb{E}_{f_X}\,\delta_{\beta_0,\pi(\gamma_*)}(X).\end{aligned} \tag{S13}$$

The minimum-MSE influence function satisfies

$$\begin{aligned}&h_\epsilon^{\mathrm{MMSE}}(\cdot,\cdot,\beta_0,\gamma_*,f_X)=\\&\underset{h(\cdot,\cdot,\beta_0,\gamma_*,f_X)}{\mathrm{argmin}}\left\{\epsilon\,\left\|\mathbb{E}_{f_X}\,\nabla_\pi\delta_{\beta_0,\pi(\gamma_*)}(X)-\mathbb{E}_{f_X}\,\mathbb{E}_{\beta_0,\pi(\gamma_*)}\,h(Y,X,\beta_0,\gamma_*,f_X)\,\nabla_\pi\log f_{\beta_0,\pi(\gamma_*)}(Y\,|\,X)\right\|_{\gamma_*}^2\right.\\&\qquad\left.+\,\frac{\mathbb{E}_{f_X}\,\mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h(Y,X,\beta_0,\gamma_*,f_X)\,|\,X)}{n}\right\}\qquad\text{subject to (S12) and (S13).}\end{aligned}$$

In the locally quadratic case, following similar derivations as for equation (18) in Lemma 1, we obtain (21).

S1.4 Corollary 1

This is a direct implication of (20).

S1.5 Corollary 2

This is a direct implication of (19).

S1.6 Corollary 3

Lemma 2 implies, analogously to (19), that

$$\begin{aligned}h_\epsilon^{\mathrm{MMSE}}(y,x)\;=\;&\delta(x)-\mathbb{E}_{f_X}\delta(X)+s_{\beta\gamma}(y\,|\,x)'[\mathbb{E}_{f_X}H_{\beta\gamma}(X)]^{-1}\,\mathbb{E}_{f_X}\nabla_{\beta\gamma}\delta(X)\\&+(\epsilon n)\widetilde{s}_\pi(y\,|\,x)^\top\left\{\mathbb{E}_{f_X}\nabla_\pi\delta(X)-\mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)s_\pi(Y\,|\,X)\right]\right\}.\end{aligned} \tag{S14}$$

Since A and X are independent, $\mathbb{E}_{f_X}\nabla_\pi\delta(X)$ can be represented by the function

$$a\mapsto\mathbb{E}_{f_X}\left[\Delta(a,X)\right]-\mathbb{E}_{f_X}\delta(X).$$

Likewise, $\mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)s_\pi(Y\,|\,X)\right]$ can be represented by the function

$$a\mapsto\mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)\,|\,A=a,X\right]=\overline{h}_\epsilon^{\mathrm{MMSE}}(a).$$

Moreover, we have for any cotangent element u (a function of a),

$$\widetilde{s}_\pi(y \mid x)^\top u = \mathbb{E}\left[u(A) \,\middle|\, Y = y, X = x\right] - \mathbb{E}\left[u(A)\right] \\ - s_{\beta\gamma}(y \mid x)'[\mathbb{E}_{f_X} H_{\beta\gamma}(X)]^{-1} \mathbb{E}_{f_X} \mathbb{E}\left[s_{\beta\gamma}(Y \mid X) u(A)\right]. \tag{S15}$$

Corollary 3 then follows from evaluating (S15) at

$$u(a) := \mathbb{E}_{f_X}\left[\Delta(a, X)\right] - \mathbb{E}_{f_X}\delta(X) - \overline{h}_\epsilon^{\text{MMSE}}(a).$$

S1.7 Corollary 4

Let us start again from (S14). In the correlated case, $\mathbb{E}_{f_X} \nabla_\pi \delta(X)$ can be represented by the function

$$(a, x) \mapsto \Delta(a, x) f_X(x) - \delta(x) f_X(x).$$

Likewise, $\mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y, X) s_\pi(Y \mid X)\right]$ can be represented by the function

$$(a, x) \mapsto \mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y, X) \mid A = a, X = x\right] f_X(x) - \mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y, X) \mid X = x\right] f_X(x) \\ = \overline{h}_\epsilon^{\text{MMSE}}(a, x) f_X(x) - \mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y, X) \mid X = x\right] f_X(x).$$

Now, by (S13) we have

$$\mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y, X) \mid X = x\right] = \delta(x) - \mathbb{E}_{f_X}\delta(X). \tag{S16}$$

Hence, $\mathbb{E}_{f_X} \nabla_\pi \delta(X) - \mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y, X) s_\pi(Y \mid X)\right]$ can be represented by the function

$$(a, x) \mapsto \Delta(a, x) f_X(x) - \mathbb{E}_{f_X}\delta(X) f_X(x) - \overline{h}_\epsilon^{\text{MMSE}}(a, x) f_X(x).$$

In the present case, cotangent elements are functions of a and x. The corresponding squared dual norm is[3]

$$\|u\|_{\gamma_*}^2 = \mathbb{E}_{f_X}\mathbb{E}\left[\left(\frac{u(A, X) - \mathbb{E}[u(A, X) \mid X]}{f_X(X)}\right)^2\right].$$

In addition we have, for any cotangent element u (a function of a and x)

$$\widetilde{s}_\pi(y \mid x)^\top u = \mathbb{E}\left[\frac{u(A, X)}{f_X(X)} \mid Y = y, X = x\right] - \mathbb{E}\left[\frac{u(A, X)}{f_X(X)} \mid X = x\right] \\ - s_{\beta\gamma}(y \mid x)'[\mathbb{E}_{f_X} H_{\beta\gamma}(X)]^{-1} \mathbb{E}_{f_X}\mathbb{E}\left[s_{\beta\gamma}(Y \mid X)\frac{u(A, X)}{f_X(X)}\right]. \tag{S17}$$

Corollary 4 then follows from evaluating (S17) at

$$u(a, x) := \Delta(a, x) f_X(x) - \mathbb{E}_{f_X}\delta(X) f_X(x) - \overline{h}_\epsilon^{\text{MMSE}}(a, x) f_X(x),$$

and noting that, by (S16), $\mathbb{E}[u(A, X) \mid X = x] = 0$.

[3]This can be shown as in Subsection S2.1, with the difference that here twice the KL divergence reads, using the notation of that subsection, $d(f_0, f_*) = -2\,\mathbb{E}_{f_X}\mathbb{E}_0 \log \frac{f_*(A \mid X)}{f_0(A \mid X)}$. Alternatively, Corollary 4 can be derived by defining π_0 as the joint distribution of (A, X), and imposing the constraint that $\int_{\mathcal{A}} \pi_0(a, x) da = f_X(x)$.

S2 Complements to Section 3

S2.1 Dual of the Kullback-Leibler divergence

Let A be a random variable with domain $\mathcal{A}$, reference distribution $f_*(a)$ and "true" distribution $f_0(a)$. We use notation $f_*(a)$ and $f_0(a)$ as if those were densities, but point masses are also allowed. Twice the Kullback-Leibler (KL) divergence reads

$$d(f_0, f_*) = -\,2\,\mathbb{E}_0 \log \frac{f_*(A)}{f_0(A)},$$

where $\mathbb{E}_0$ is the expectation under f_0. Let $\mathcal{F}$ be the set of all distributions, in particular, $f \in \mathcal{F}$ implies $\int_{\mathcal{A}} f(a)da = 1$. Let $q : \mathcal{A} \to \mathbb{R}$ be a real valued function. For given $f_* \in \mathcal{F}$ and $\epsilon > 0$ we define

$$\|q\|_{*,\epsilon} := \max_{\{f_0 \in \mathcal{F} \,:\, d(f_0,f_*) \le \epsilon\}} \frac{\mathbb{E}_0\, q(A) - \mathbb{E}_*\, q(A)}{\sqrt{\epsilon}},$$

where $\mathbb{E}_*$ is the expectation under f_*.

We have the following result.

Lemma S2. *For $q : \mathcal{A} \to \mathbb{R}$ and $f_* \in \mathcal{F}$ we assume that the moment-generating function $m_*(t) = \mathbb{E}_* \exp(t\, q(A))$ exists for $t \in (\delta_-, \delta_+)$ and some $\delta_- < 0$ and $\delta_+ > 0$.*[4] *For $\epsilon \in (0, \delta_+^2)$ we then have*

$$\|q\|_{*,\epsilon} = \sqrt{\mathrm{Var}_*(q(A))} + O(\epsilon^{\frac{1}{2}}).$$

Proof. Let the cumulant-generating function of the random variable $q(A)$ under the reference measure f_* be $k_*(t) = \log m_*(t)$. We assume existence of $m_*(t)$ and $k_*(t)$ for $t \in (\delta_-, \delta_+)$. This also implies that all derivatives of $m_*(t)$ and $k_*(t)$ exist in this interval. We denote the p-th derivative of $m_*(t)$ by $m_*^{(p)}(t)$, and analogously for $k_*(t)$.

In the following we denote the maximizing f_0 in the definition of $\|q\|_{*,\epsilon}$ simply by f_0. Applying standard optimization method (Karush-Kuhn-Tucker) we find the well-known exponential tilting result

$$f_0(a) = c\, f_*(a)\, \exp(t\, q(a)),$$

where the constants $c, t \in (0, \infty)$ are determined by the constraints $\int_{\mathcal{A}} f_0(a)da = 1$ and $d(f_0, f_*) = \epsilon$. Using the constraint $\int_{\mathcal{A}} f_0(a)da = 1$ we can solve for c to obtain

$$f_0(a) = \frac{f_*(a)\, \exp(t\, q(a))}{\mathbb{E}_* \exp(t\, q(A))} = \frac{f_*(a)\, \exp(t\, q(a))}{m_*(t)}.$$

[4]Existence of $m_*(t)$ in an open interval around zero is equivalent to having an exponential decay of the tails of the distribution of the random variable $Q = q(A)$. If $q(a)$ is bounded, then $m_*(t)$ exists for all $t \in \mathbb{R}$.

Using this we find that

$$\begin{aligned}
d(t) &:= d(f_0, f_*) \\
&= 2\,\mathbb{E}_* \frac{f_0(A)}{f_*(A)} \log \frac{f_0(A)}{f_*(A)} \\
&= \frac{2\,t}{m_*(t)}\,\mathbb{E}_* \exp(t\,q(A))q(A) - \frac{2\log m_*(t)}{m_*(t)}\,\mathbb{E}_* \exp(t\,q(A)) \\
&= \frac{2\,t\,m_*^{(1)}(t)}{m_*(t)} - 2\log m_*(t) \\
&= 2\left[t\,k_*^{(1)}(t) - k_*(t)\right].
\end{aligned}$$

We have $d(0) = 0$, $d^{(1)}(0) = 0$, $d^{(2)}(0) = 2k_*^{(2)}(0) = 2\mathrm{Var}_*(q(A))$, $d^{(3)}(t) = 4k_*^{(3)}(t) + 2tk_*^{(4)}(t)$. A mean-value expansion thus gives

$$d(t) = \mathrm{Var}_*(q(A))t^2 + \frac{t^3}{6}\left[4\,k_*^{(3)}(\tilde{t}) + 2\,\tilde{t}\,k_*^{(4)}(\tilde{t})\right],$$

where $0 \le \tilde{t} \le t \le \delta_+$. The value t that satisfies the constraint $d(t) = \epsilon$ therefore satisfies

$$t = \frac{\epsilon^{\frac{1}{2}}}{\sqrt{\mathrm{Var}_*(q(A))}} + O(\epsilon).$$

Next, using that $\|q\|_{*,\epsilon} = \epsilon^{-\frac{1}{2}}\,\mathbb{E}_*\left[\left(\frac{f_0(A)}{f_*(A)} - 1\right) q(A)\right]$ we find

$$\|q\|_{*,\epsilon} = \epsilon^{-\frac{1}{2}}\left[k_*^{(1)}(t) - k_*^{(1)}(0)\right].$$

Again using that $k_*^{(2)}(0) = \mathrm{Var}_*(q(A))$ and applying a mean value expansion we obtain

$$\begin{aligned}
\|q\|_{*,\epsilon} &= \epsilon^{-\frac{1}{2}}\left[t\,k_*^{(2)}(t) + \frac{1}{2}\,t^2\,k_*^{(3)}(\bar{t})\right] \\
&= \epsilon^{-\frac{1}{2}}\left[t\,\mathrm{Var}_*(q(A)) + \frac{1}{2}\,t^2\,k_*^{(3)}(\bar{t})\right] \\
&= \sqrt{\mathrm{Var}_*(q(A))} + O(\epsilon^{\frac{1}{2}}),
\end{aligned}$$

where $\bar{t} \in [0, t]$. ∎

S2.2 Equations (25), (26) and (27)

Here we use simplified notation as in Section 3. Let us start by deriving (25). In this case β_0 and γ_* are known, and Corollary 2 gives

$$h_\epsilon^{\mathrm{MMSE}} = (\epsilon n)\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\left[\Delta - \delta - \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} h^{\mathrm{MMSE}}\right],$$

so

$$h_\epsilon^{\mathrm{MMSE}} = \left[(\epsilon n)^{-1}\mathbb{I}_{\mathcal{Y}} + \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \circ \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}\right]^{-1} \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\left[\Delta - \delta\right].$$

(25) then follows from the operator identity

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

Let us now derive (26). In this case γ_* is known. Since $\Delta(A) = c'\beta_0 = \delta$, Corollary 2 implies

$$h_\epsilon^{\mathrm{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} c - (\epsilon n)\left\{\mathbb{E}\left[\overline{h}^{\mathrm{MMSE}}(A)\,|\,Y = y\right] - s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \mathbb{E}\left[s_{\beta\gamma}(Y)\overline{h}^{\mathrm{MMSE}}(A)\right]\right\}.$$

Hence, we have, for some vector b,

$$h_\epsilon^{\mathrm{MMSE}} = s_{\beta\gamma}(y)'b - (\epsilon n)\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \circ \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} h^{\mathrm{MMSE}}.$$

Using the Woodbury identity

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

we thus obtain

$$h_\epsilon^{\mathrm{MMSE}} = \mathbb{W}^\epsilon s_{\beta\gamma}(y)'b.$$

Lastly, since by (4) $\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y)s_{\beta\gamma}(Y)\right] = c$, we obtain (26) whenever the denominator is non-singular.

Finally, let us derive (27). In this case β_0 and γ_* are known and $\Delta(A)$ does not depend on X, and Corollary 3 gives

$$h_\epsilon^{\mathrm{MMSE}} = (\epsilon n)\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}} \left[\mathbb{E}_{f_X}(\Delta - \delta) - \mathbb{E}_{\mathcal{Y},\mathcal{X}\,|\,\mathcal{A}} h^{\mathrm{MMSE}}\right].$$

Hence, denoting $\mathbb{I}_{\mathcal{Y},\mathcal{X}} h(y,x) = h(y,x)$ the identity operator, we have

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

(27) then follows from

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

S3 Computation in semi-parametric models

Here we describe how we compute a numerical approximation to the minimum-MSE estimator in semi-parametric models

$$\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \mathbb{E}_{\widehat{\beta},\pi(\widehat{\gamma})}\,\Delta_{\widehat{\beta}}(A) + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i, \widehat{\beta}, \widehat{\gamma}),$$

where $h_\epsilon^{\mathrm{MMSE}}$ is given by Corollary 2, and $\widehat{\beta}, \widehat{\gamma}$ are preliminary estimates. We abstract from conditioning covariates. In the presence of correlated covariates X_i we use the same technique to approximate $h_\epsilon^{\mathrm{MMSE}}(\cdot\,|\,x)$ for each value of $X_i = x$. We use this approach in the numerical illustration based on the dynamic panel data model in Section 5, where the covariate is the initial condition. We denote $\eta = (\beta', \gamma')'$.[5]

Draw an i.i.d. sample $(Y^{(1)}, A^{(1)}), ..., (Y^{(S)}, A^{(S)})$ of S draws from $g_\beta \times \pi(\gamma)$. Let G be $S \times S$ with (τ, s) element $g_\beta(Y^{(\tau)} \,|\, A^{(s)}) / \sum_{s'=1}^S g_\beta(Y^{(\tau)} \,|\, A^{(s')})$, G_Y be $N \times S$ with (i, s) element $g_\beta(Y_i \,|\, A^{(s)}) / \sum_{s'=1}^S g_\beta(Y_i \,|\, A^{(s')})$, Δ be $S \times 1$ with s-th element $\Delta_\beta(A^{(s)})$, I be the $S \times S$ identity matrix, and ι and ι_Y be the $S \times 1$ and $N \times 1$ vectors of ones. In addition, let D be the $S \times \dim \eta$ matrix with (s, k) element

$$d_{\eta_k}(Y^{(s)}) = \frac{\sum_{s'=1}^S \left(\nabla_{\eta_k} \log g_\beta(Y^{(s)} \,|\, A^{(s')}) + \nabla_{\eta_k} \log \pi(\gamma)(A^{(s')}) \right) g_\beta(Y^{(s)} \,|\, A^{(s')})}{\sum_{s'=1}^S g_\beta(Y^{(s)} \,|\, A^{(s')})},$$

and let D_Y be $N \times \dim \eta$ with (i, k) element $d_{\eta_k}(Y_i)$, $Q = I - DD^\dagger$, $\widetilde{G}_Y = G_Y - D_Y D^\dagger G$, $\widetilde{\iota}_Y = \iota_Y - D_Y D^\dagger \iota$, $\widetilde{G} = QG$, $\widetilde{\iota} = Q\iota$, and $\partial\Delta$ be the $K \times 1$ vector with k-th element $\frac{1}{S} \sum_{s=1}^S \nabla_{\eta_k} \Delta(A^{(s)}, \beta) + \Delta(A^{(s)}, \beta) \nabla_{\eta_k} \log \pi(\gamma)(A^{(s)})$.

From Corollary 2, a fixed-S approximation to the minimum-MSE estimator is then

$$\widetilde{\delta}_\epsilon^{\text{MMSE}} = \iota^\dagger \Delta + \iota_Y^\dagger \widetilde{h}_\epsilon^{\text{MMSE}},$$

where

$$\widetilde{h}_\epsilon^{\text{MMSE}} = D_Y (D'D/S)^{-1} \partial\Delta + (\epsilon n) \Bigg[\left(\widetilde{G}_Y - \widetilde{\iota}_Y \iota^\dagger \right) \Delta - \widetilde{G}_Y G' \left(\widetilde{G} G' + (\epsilon n)^{-1} I \right)^{-1} \left((\epsilon n)^{-1} D (D'D/S)^{-1} \partial\Delta + \left(\widetilde{G} - \widetilde{\iota} \iota^\dagger \right) \Delta \right) \Bigg],$$

and (β, γ) are replaced by the preliminary $(\widehat{\beta}, \widehat{\gamma})$ in all the quantities above, including when producing the simulated draws.

Confidence intervals. From Subsection 2.4, computing confidence intervals only requires, in addition to computing critical values under correct specification, to compute an estimate of the bias of the estimator $b_\epsilon(h, \widehat{\beta}, \widehat{\gamma})$. In semi-parametric models we have, for an asymptotically linear estimator based on h satisfying (2) and (4),

$$b_\epsilon(h, \beta_0, \gamma_*) = \epsilon^{\frac{1}{2}} \left\{ \text{Var}_{\beta_0, \pi(\gamma_*)} [\Delta_{\beta_0}(A) - \mathbb{E}_{\beta_0, \pi(\gamma_*)}(h(Y) \,|\, A)] \right\}^{\frac{1}{2}}.$$

A numerical approximation of the bias of $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is then

$$\widetilde{b}_\epsilon(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) = \epsilon^{\frac{1}{2}} \left\| \Delta - \iota^\dagger \Delta - G' \widetilde{h}_\epsilon^{\text{MMSE}} \right\|.$$

Values of ϵ. In turn, ϵ_k in (31) can be approximated as $\mu(\alpha, p)^2 / (n\lambda_k)$, where λ_k is the k-th largest eigenvalue of $G'QG = \widetilde{G}'\widetilde{G}$ (removing the eigenvalue equal to one since it corresponds to a constant eigenfunction). We proceed similarly to compute eigenfunctions in Subsection 5.1.

[5] Here we present a general method based on simulations. In the cross-sectional probit model (32), explicit closed-form expressions are available, and we use those for computation in our first illustration.

S4 Application to structural evaluation of 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. Our approach provides a way to improve the policy predictions of a structural model when the model may be misspecified. It does not require the researcher to estimate another (larger) structural model, and provides a tractable way to perform sensitivity analysis in such settings.

S4.1 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. Moreover, 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 (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. Note the direct presence 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 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 (that is, of the subsidy τ) we obtain

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

where Φ is the standard normal cdf.

We estimate the model on control villages, under the assumption that $\lambda = 0$. The average effect of the subsidy on school attendance is

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

Note that 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 by design. TW use a similar strategy to predict the effect of the program and other counterfactual policies, in the spirit of "ex-ante" policy prediction. Here we use the specification with $\lambda = 0$ as our reference model.

As Wolpin (2013) notes, in the presence of a stigma effect (i.e., when $\lambda \neq 0$) information from treated villages is needed for identification and estimation.[6] Instead of estimating a larger model, here we adjust the predictions from the reference model against the possibility of misspecification, using data from both controls and treated. While in the present simple static context one could easily estimate a version of the model allowing for $\lambda \neq 0$, 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

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

We focus on the effect on eligible (i.e., poorer) households. We will first estimate $\delta_{\beta,0}$ using the control villages only. We will then compute our minimum-MSE estimator, taking advantage of the variation in treatment status in order to account for the potential misspecification. We will also report confidence intervals. In this setting our assumption that ϵ shrinks as n increases reflects that the econometrician's uncertainty about the presence of stigma effects diminishes when the sample gets larger.

S4.2 Empirical results

We use the sample from TW. We drop observations with missing household income, and focus on boys and girls aged 12 to 15. This results in 1219 (boys) and 1089 (girls) observations, respectively. 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. 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 age and gender. Average school attendance drops from 90% at age 12 to between 40% and 50% at age 15.

[6] 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.

In Table S5 we show the results of different estimators and confidence intervals. The top panel focuses on the impact of the PROGRESA subsidy on eligible households. The left two columns show the point estimates of the policy impact as well as 95% confidence intervals, calculated under the assumption that the reference model is correct (second row) and under the assumption that the model belongs to an ϵ-neighborhood of the reference model (third row). We show the results for our focal value $\epsilon = \epsilon_1$. The model-based predictions are calculated based on control villages. We add covariates to the gender-specific school attendance equations, which include the age of the child and her parents, year indicators, distance to school, and an eligibility indicator. In the middle two columns of Table S5 we report estimates of the minimum-MSE estimator for the same ϵ, together with confidence intervals. The minimum-MSE estimates are computed based on both treated and control villages. Lastly, in the right two columns we report the differences in means between treated and control villages.

We see that PROGRESA had 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) are very large for both genders. This suggests that model misspecification, such as the presence of a stigma effect of the program, may strongly affect the ability to produce "ex-ante" policy predictions in this context. When adding treated villages to the sample and computing our minimum-MSE estimators, we find that the effect for girls is similar to the baseline specification, whereas the effect for boys is smaller, around 5 percentage points. Moreover, the length of the confidence intervals is then substantially reduced, although they are still large. Interestingly, as shown by the rightmost two columns the minimum-MSE estimates are quite close to the experimental differences in means between treated and control villages, for both genders.

When using our approach it is informative to report minimum-MSE estimates and confidence intervals for different values of the neighborhood size ϵ. In Figure S1 we plot the estimates for girls (left) and boys (right) as a function of ϵ, in addition to 95% confidence intervals based on those estimates. In dotted we show the unadjusted model-based predictions. The minimum-MSE estimates vary very little with ϵ for girls, and show slightly more variation for boys. Note that the minimum-MSE estimate at $\epsilon = 0$ for boys is .058, compared to .054 for our focal ϵ value, and .080 for the estimate predicted by the reference model estimated on control villages. This suggests that, for boys, the functional form of the schooling decision is *not* invariant to treatment status, again highlighting that predictions based off the controls are less satisfactory for boys (as acknowledged by Todd and Wolpin, 2006).

On the middle and bottom panels of Table S5 we next show estimates, based on the reference model and minimum-MSE adjustments, of the effects of two counterfactual policies: doubling the PROGRESA subsidy, and removing the conditioning of the income transfer on school attendance. Unlike in the case of the main PROGRESA effects, there is no experimental counterpart to such counterfactuals. Estimates based on our approach predict a substantial effect of doubling the subsidy on girls' attendance and a more moderate effect on boys. By contrast, we find no effect of an unconditional income transfer.

Given the structural model, one can give an economic interpretation to the value of ϵ.

To see this, note that in the model λ is the marginal utility of the subsidy for households sending their child to school. The marginal utility of consumption is $a+d$. Hence, bounding λ^2 by ϵ is equivalent to bounding the ratio of marginal utility of the subsidy to marginal utility of consumption by $\sqrt{\epsilon}/(a+d)$. Taking $\epsilon = \epsilon_1$ implies that this ratio of marginal utilities is bounded by 2.2 (females) and 1.6 (males).

Lastly, the analysis in this section is based on a reference model estimated on the subsample of control villages, as in TW. Treated villages are only added when constructing minimum-MSE estimators. An alternative approach, in the spirit of "ex-post" policy prediction, is to estimate the reference model on both controls and treated, and perform the adjustments based on the same data. The results are similar to Table S5 and Figure S1, except for the fact that model-based, minimum-MSE and experimental estimates are more similar for boys in that case (results are available upon request).

S5 Models defined by moment restrictions

In this section we consider settings where a finite-dimensional parameter $(\beta_0', \pi_0')'$ does not fully determine the distribution f_0 of Y, but satisfies a finite-dimensional system of moment conditions

$$\mathbb{E}_{f_0} \Psi(Y, \beta_0, \pi_0) = 0, \tag{S18}$$

which may be just-identified, over-identified or under-identified. We focus on asymptotically linear generalized method-of-moments (GMM) estimators of δ_{β_0, π_0} that satisfy

$$\widehat{\delta} = \delta_{\beta_0, \pi(\gamma_*)} + a(\beta_0, \gamma_*)' \frac{1}{n} \sum_{i=1}^{n} \Psi(Y_i, \beta_0, \pi(\gamma_*)) + o_{P_0}(\epsilon^{\frac{1}{2}} + n^{-\frac{1}{2}}), \tag{S19}$$

for a parameter vector $a(\beta_0, \gamma_*)$. We will characterize the form of $a(\beta_0, \gamma_*)$ leading to minimum worst-case MSE in $\Gamma_\epsilon(\gamma_*)$.

We assume that the remainder in (S19) is uniformly bounded similarly as in (14). In this case local robustness with respect to $(\beta_0', \gamma_*')'$ takes the form

$$\nabla_{\beta\gamma} \delta_{\beta_0, \pi(\gamma_*)} + \mathbb{E}_{f_0} \nabla_{\beta\gamma} \Psi(Y, \beta_0, \pi(\gamma_*))\, a(\beta_0, \gamma_*) = 0. \tag{S20}$$

It is natural to focus on asymptotically linear GMM estimators here, since f_0 is unrestricted except for the moment condition (S18).

To derive the worst-case bias of $\widehat{\delta}$ note that, by (S18), for any $\pi_0 \in \Gamma_\epsilon(\gamma_*)$ we have

$$\mathbb{E}_{f_0} \Psi(Y, \beta_0, \pi(\gamma_*)) = - \left[\mathbb{E}_{f_0} \nabla_\pi \Psi(Y, \beta_0, \pi(\gamma_*))\right]' (\pi_0 - \pi(\gamma_*)) + o(\epsilon^{\frac{1}{2}}),$$

so, under appropriate regularity conditions,

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \mathbb{E}_{f_0} \widehat{\delta} - \delta_{\beta_0, \pi_0} \right| = \epsilon^{\frac{1}{2}} \left\| \nabla_\pi \delta_{\beta_0, \pi(\gamma_*)} + \mathbb{E}_{f_0} \nabla_\pi \Psi(Y, \beta_0, \pi(\gamma_*))\, a(\beta_0, \gamma_*) \right\|_{\gamma_*} + o(\epsilon^{\frac{1}{2}} + n^{-\frac{1}{2}}).$$

The worst-case MSE of

$$\widehat{\delta}_{a, \beta_0, \gamma_*} := \delta_{\beta_0, \pi(\gamma_*)} + a(\beta_0, \gamma_*)' \frac{1}{n} \sum_{i=1}^{n} \Psi(Y_i, \beta_0, \pi(\gamma_*))$$

is thus

$$\epsilon \left\| \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{f_0} \nabla_\pi \Psi(Y, \beta_0, \pi(\gamma_*))\, a(\beta_0, \gamma_*) \right\|^2_{\gamma_*}$$
$$+ a(\beta_0, \gamma_*)' \frac{\mathbb{E}_{f_0} \Psi(Y, \beta_0, \pi(\gamma_*))\Psi(Y, \beta_0, \pi(\gamma_*))'}{n} a(\beta_0, \gamma_*) + o(\epsilon + n^{-1}).$$

To obtain an explicit expression for the minimum-MSE estimator, let us focus on the case where π_0 is finite-dimensional and $\|\cdot\|_{\gamma_*} = \|\cdot\|_{\Omega^{-1}}$. Let us define

$$V_{\beta_0,\pi(\gamma_*)} = \mathbb{E}_{f_0} \Psi(Y, \beta_0, \pi(\gamma_*))\Psi(Y, \beta_0, \pi(\gamma_*))', \quad K_{\beta_0,\pi(\gamma_*)} = \mathbb{E}_{f_0} \nabla_\pi \Psi(Y, \beta_0, \pi(\gamma_*)),$$

and

$$K_{\beta_0,\gamma_*} = \mathbb{E}_{f_0} \nabla_{\beta\gamma} \Psi(Y, \beta_0, \pi(\gamma_*)).$$

For all β_0, γ_* we aim to minimize

$$\epsilon \left\| \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} + K_{\beta_0,\pi(\gamma_*)} a(\beta_0, \gamma_*) \right\|^2_{\Omega^{-1}} + a(\beta_0, \gamma_*)' \frac{V_{\beta_0,\pi(\gamma_*)}}{n} a(\beta_0, \gamma_*),$$
$$\text{subject to } \nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)} + K_{\beta_0,\gamma_*} a(\beta_0, \gamma_*) = 0.$$

A solution is given by[7]

$$a_\epsilon^{\mathrm{MMSE}}(\beta_0, \gamma_*) = -B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon} K'_{\beta_0,\gamma_*} \left(K_{\beta_0,\gamma_*} B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon} K'_{\beta_0,\gamma_*} \right)^{-1} \nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)}$$
$$- B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon} \left(I - K'_{\beta_0,\gamma_*} \left(K_{\beta_0,\gamma_*} B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon} K'_{\beta_0,\gamma_*} \right)^{-1} K_{\beta_0,\gamma_*} B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon} \right) K'_{\beta_0,\pi(\gamma_*)} \Omega^{-1} \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)}, \tag{S21}$$

where $B_{\beta_0,\pi(\gamma_*),\epsilon} = K'_{\beta_0,\pi(\gamma_*)} \Omega^{-1} K_{\beta_0,\pi(\gamma_*)} + (\epsilon n)^{-1} V_{\beta_0,\pi(\gamma_*)}$, and $B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon}$ is its Moore-Penrose generalized inverse. Note that, in the likelihood case and taking $\Psi(y, \beta, \pi) = \nabla_\pi \log f_{\beta,\pi}(y)$, the function $h(y, \beta_0, \gamma_*) = a_\epsilon^{\mathrm{MMSE}}(\beta_0, \gamma_*)'\Psi(y, \beta_0, \pi(\gamma_*))$ simplifies to (20).

As a special case, when $\epsilon = 0$ we have

$$a_0^{\mathrm{MMSE}}(\beta_0, \gamma_*) = -V^\dagger_{\beta_0,\pi(\gamma_*)} K'_{\beta_0,\gamma_*} \left(K_{\beta_0,\gamma_*} V^\dagger_{\beta_0,\pi(\gamma_*)} K'_{\beta_0,\gamma_*} \right)^{-1} \nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)}.$$

In this case, given preliminary estimators $\widehat{\beta}$ and $\widehat{\gamma}$, the minimum-MSE estimator

$$\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \delta_{\widehat{\beta},\pi(\widehat{\gamma})} + a_0^{\mathrm{MMSE}}(\widehat{\beta}, \widehat{\gamma})' \frac{1}{n} \sum_{i=1}^n \Psi(Y_i, \widehat{\beta}, \pi(\widehat{\gamma}))$$

is the one-step approximation to the optimal GMM estimator based on the reference model. To obtain a feasible estimator one simply replaces the expectations in $V_{\beta_0,\pi(\gamma_*)}$ and K_{β_0,γ_*} by sample analogs.

[7] Here we assume that $K_{\beta_0,\gamma_*} V^\dagger_{\beta_0,\pi(\gamma_*)} K'_{\beta_0,\gamma_*}$ is non-singular, requiring that β_0, γ_* be identified from the moment conditions. Existence follows from the fact that, by the generalized information identity, $V_{\beta_0,\pi(\gamma_*)} a = 0$ implies that $K_{\beta_0,\pi(\gamma_*)} a = 0$. Moreover, although $a_\epsilon^{\mathrm{MMSE}}(\beta_0, \gamma_*)$ may not be unique, $a_\epsilon^{\mathrm{MMSE}}(\beta_0, \gamma_*)'\Psi(Y, \beta_0, \pi(\gamma_*))$ is unique almost surely.

As a second special case, consider ϵ tending to infinity. Focusing on the known-(β_0, γ_*) case for simplicity, $a_\epsilon^{\text{MMSE}}(\beta_0, \gamma_*)$ tends to $-K^{\text{ginv}}_{\beta_0,\pi(\gamma_*)} \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)}$, where

$$K^{\text{ginv}}_{\beta_0,\pi(\gamma_*)} :=$$
$$\left(V^\dagger_{\beta_0,\pi(\gamma_*)}\right)^{1/2} \left[\left(V^\dagger_{\beta_0,\pi(\gamma_*)}\right)^{1/2} K'_{\beta_0,\pi(\gamma_*)}\Omega^{-1}K_{\beta_0,\pi(\gamma_*)} \left(V^\dagger_{\beta_0,\pi(\gamma_*)}\right)^{1/2}\right]^\dagger \left(V^\dagger_{\beta_0,\pi(\gamma_*)}\right)^{1/2} K'_{\beta_0,\pi(\gamma_*)}\Omega^{-1}$$

is a generalized inverse of $K_{\beta_0,\pi(\gamma_*)}$, and the choice of Ω corresponds to choosing one specific such generalized inverse. In this case, the minimum-MSE estimator is the one-step approximation to a particular GMM estimator based on the "large" model.

Lastly, given a parameter vector a, confidence intervals can be constructed as explained in Subsection 2.4, taking

$$b_\epsilon(a, \widehat{\beta}, \widehat{\gamma}) = \epsilon^{\frac{1}{2}} \left\| \nabla_\pi \delta_{\widehat{\beta},\pi(\widehat{\gamma})} + \frac{1}{n}\sum_{i=1}^n \nabla_\pi \Psi(Y_i, \widehat{\beta}, \pi(\widehat{\gamma}))\, a(\widehat{\beta}, \widehat{\gamma}) \right\|_{\Omega^{-1}}.$$

Example. Consider again the OLS/IV example of Subsection 3.3, but now drop the Gaussian assumptions on the distributions. For known C, the set of moment conditions corresponds to the moment functions

$$\Psi(y, x, z, \beta, \pi) = \begin{pmatrix} x(y - x'\beta - \pi'(x - Cz)) \\ z(y - x'\beta) \end{pmatrix}.$$

In this case, letting $W = (X', Z')'$ we have

$$K_{\beta_0,\gamma_*} = -\mathbb{E}_{f_0}\left(XW'\right), \quad K_{\beta_0,\pi(\gamma_*)} = -\mathbb{E}_{f_0}\begin{pmatrix} XX' & XZ' \\ (X - CZ)X' & 0 \end{pmatrix},$$

and

$$V_{\beta_0,\pi(\gamma_*)} = \mathbb{E}_{f_0}\left((Y - X'\beta_0)^2 WW'\right).$$

Given a preliminary estimator $\widetilde{\beta}$, $V_{\beta_0,\pi(\gamma_*)}$ can be estimated as $\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widetilde{\beta})^2 W_i W_i'$, whereas K_{β_0,γ_*} and $K_{\beta_0,\pi(\gamma_*)}$ can be estimated as sample means. The estimator based on (S21) then interpolates nonlinearly between the OLS and IV estimators, similarly as in the likelihood case.

S6 Additional simulation results

Table S1: Monte Carlo simulation of the average effect in the cross-sectional binary choice model, interpolation ($x_0 = (.5, 1)'$)

Minimum-MSE, for $\epsilon =$	0.00	0.20	0.40	0.60	0.80	1.00
	A. $n_X = 4$					
Worst-case bias	0.0021	0.0783	0.1104	0.1351	0.1560	0.1744
Asymptotic standard error	0.0228	0.0288	0.0297	0.0300	0.0302	0.0303
Monte Carlo bias	0.1026	0.0197	0.0134	0.0111	0.0099	0.0092
Monte Carlo standard deviation	0.0253	0.0281	0.0288	0.0291	0.0292	0.0293
Monte Carlo root MSE	0.1057	0.0343	0.0317	0.0311	0.0308	0.0307
CI length	0.0936	0.2697	0.3372	0.3878	0.4302	0.4674
CI coverage	0.0180	0.9990	1.0000	1.0000	1.0000	1.0000
	B. $n_X = 20$					
Worst-case bias	0.0021	0.0480	0.0610	0.0714	0.0805	0.0887
Asymptotic standard error	0.0227	0.0394	0.0453	0.0487	0.0509	0.0526
Monte Carlo bias	0.0976	0.0080	0.0037	0.0026	0.0022	0.0020
Monte Carlo standard deviation	0.0239	0.0386	0.0446	0.0480	0.0502	0.0519
Monte Carlo root MSE	0.1005	0.0394	0.0447	0.0480	0.0502	0.0519
CI length	0.0931	0.2503	0.2996	0.3337	0.3607	0.3835
CI coverage	0.0190	0.9990	1.0000	1.0000	1.0000	1.0000

Notes: Performance of the minimum-MSE estimator in the cross-sectional binary choice model, for different values of ϵ. $n = 500$, results for 1000 simulations. The nominal level for confidence intervals (CI) is 95%. n_X denotes the number of points of support of the first component of X.

Table S2: Monte Carlo simulation of the average effect in the cross-sectional binary choice model, extrapolation ($x_0 = (-.5, 1)'$)

Minimum-MSE, for $\epsilon =$	0.00	0.20	0.40	0.60	0.80	1.00
			A. $n_X = 4$			
Worst-case bias	0.0029	0.1269	0.1794	0.2197	0.2537	0.2837
Asymptotic standard error	0.0296	0.0312	0.0315	0.0316	0.0316	0.0317
Monte Carlo bias	-0.0987	-0.0903	-0.0901	-0.0900	-0.0900	-0.0900
Monte Carlo standard deviation	0.0283	0.0330	0.0334	0.0335	0.0336	0.0336
Monte Carlo root MSE	0.1027	0.0961	0.0961	0.0961	0.0961	0.0961
CI length	0.1219	0.3762	0.4822	0.5632	0.6314	0.6914
CI coverage	0.2000	0.9370	0.9850	0.9960	0.9990	1.0000
			B. $n_X = 20$			
Worst-case bias	0.0028	0.1172	0.1645	0.2008	0.2314	0.2584
Asymptotic standard error	0.0313	0.0401	0.0443	0.0470	0.0489	0.0503
Monte Carlo bias	-0.0902	-0.0961	-0.0988	-0.0999	-0.1005	-0.1009
Monte Carlo standard deviation	0.0287	0.0373	0.0412	0.0437	0.0456	0.0471
Monte Carlo root MSE	0.0947	0.1031	0.1070	0.1090	0.1104	0.1113
CI length	0.1284	0.3915	0.5026	0.5857	0.6544	0.7141
CI coverage	0.2530	0.9500	0.9910	0.9960	0.9970	0.9970

Notes: See the notes to Table S1.

Table S3: Monte Carlo simulation results for the autoregressive parameter in the dynamic binary choice panel data model

Minimum-MSE, for $\epsilon =$	0.00	0.20	0.40	0.60	0.80	1.00
	A. $T = 5$					
Worst-case bias	0.0001	0.0337	0.0470	0.0573	0.0660	0.0737
Asymptotic standard error	0.1343	0.1378	0.1384	0.1387	0.1390	0.1392
Monte Carlo bias	-0.1729	-0.0615	-0.0555	-0.0531	-0.0518	-0.0509
Monte Carlo standard deviation	0.1252	0.1111	0.1129	0.1136	0.1141	0.1145
Monte Carlo root MSE	0.2135	0.1270	0.1258	0.1255	0.1254	0.1253
CI length	0.5268	0.6077	0.6363	0.6583	0.6768	0.6931
CI coverage	0.7610	0.9860	0.9910	0.9910	0.9930	0.9930
	B. $T = 10$					
Worst-case bias	0.0001	0.0217	0.0305	0.0373	0.0430	0.0481
Asymptotic standard error	0.0857	0.0868	0.0869	0.0870	0.0871	0.0872
Monte Carlo bias	-0.0780	-0.0137	-0.0120	-0.0114	-0.0110	-0.0107
Monte Carlo standard deviation	0.0676	0.0731	0.0736	0.0738	0.0739	0.0740
Monte Carlo root MSE	0.1032	0.0744	0.0745	0.0746	0.0747	0.0748
CI length	0.3360	0.3835	0.4017	0.4156	0.4274	0.4378
CI coverage	0.9090	0.9900	0.9910	0.9920	0.9930	0.9940
	C. $T = 20$					
Worst-case bias	0.0001	0.0164	0.0231	0.0283	0.0327	0.0366
Asymptotic standard error	0.0590	0.0596	0.0596	0.0597	0.0597	0.0598
Monte Carlo bias	-0.0304	-0.0023	-0.0019	-0.0017	-0.0017	-0.0016
Monte Carlo standard deviation	0.0442	0.0488	0.0490	0.0490	0.0491	0.0491
Monte Carlo root MSE	0.0537	0.0488	0.0490	0.0491	0.0491	0.0491
CI length	0.2316	0.2663	0.2800	0.2906	0.2995	0.3074
CI coverage	0.9720	0.9950	0.9960	0.9970	0.9980	0.9990

Notes: Performance of the minimum-MSE estimator of β_0 in the dynamic panel data binary choice model, for different values of ϵ. $n = 500$, results for 1000 simulations. The nominal level for confidence intervals (CI) is 95%.

Table S4: Monte Carlo simulation results for the average state dependence parameter in the dynamic binary choice panel data model

Minimum-MSE, for $\epsilon =$	0.00	0.20	0.40	0.60	0.80	1.00
	A. $T = 5$					
Worst-case bias	0.0001	0.0260	0.0360	0.0437	0.0501	0.0558
Asymptotic standard error	0.0373	0.0386	0.0388	0.0390	0.0392	0.0393
Monte Carlo bias	-0.0538	-0.0218	-0.0202	-0.0196	-0.0193	-0.0191
Monte Carlo standard deviation	0.0439	0.0324	0.0331	0.0334	0.0336	0.0337
Monte Carlo root MSE	0.0694	0.0391	0.0387	0.0387	0.0387	0.0388
CI length	0.1463	0.2033	0.2243	0.2403	0.2538	0.2657
CI coverage	0.6380	0.9830	0.9900	0.9960	0.9970	0.9980
	B. $T = 10$					
Worst-case bias	0.0001	0.0356	0.0499	0.0609	0.0701	0.0782
Asymptotic standard error	0.0266	0.0269	0.0271	0.0272	0.0272	0.0273
Monte Carlo bias	-0.0212	-0.0047	-0.0048	-0.0050	-0.0051	-0.0052
Monte Carlo standard deviation	0.0257	0.0229	0.0230	0.0231	0.0231	0.0232
Monte Carlo root MSE	0.0333	0.0233	0.0235	0.0236	0.0237	0.0238
CI length	0.1045	0.1769	0.2059	0.2282	0.2469	0.2634
CI coverage	0.8660	1.0000	1.0000	1.0000	1.0000	1.0000
	C. $T = 20$					
Worst-case bias	0.0001	0.0442	0.0622	0.0761	0.0878	0.0980
Asymptotic standard error	0.0198	0.0200	0.0200	0.0201	0.0202	0.0202
Monte Carlo bias	-0.0097	-0.0028	-0.0028	-0.0028	-0.0028	-0.0028
Monte Carlo standard deviation	0.0187	0.0153	0.0153	0.0154	0.0154	0.0155
Monte Carlo root MSE	0.0210	0.0155	0.0156	0.0156	0.0157	0.0157
CI length	0.0777	0.1666	0.2030	0.2310	0.2546	0.2754
CI coverage	0.9290	1.0000	1.0000	1.0000	1.0000	1.0000

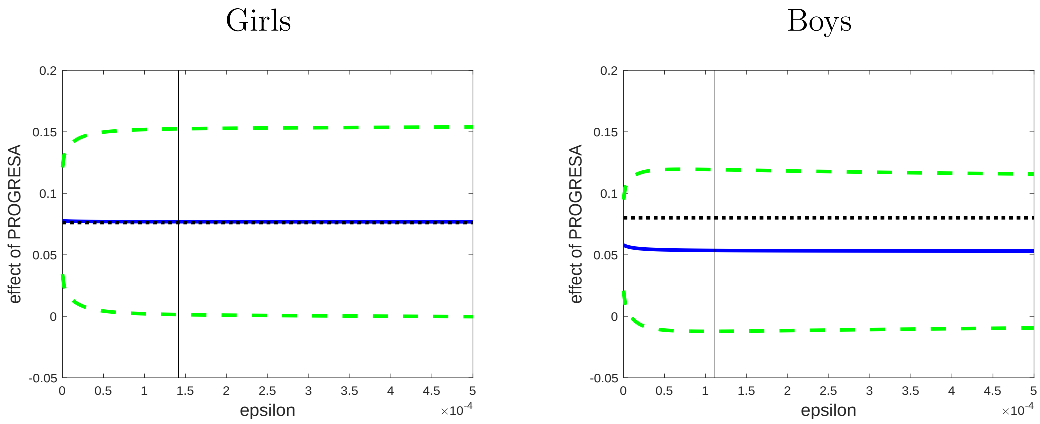
Notes: Performance of the minimum-MSE estimator of δ_{β_0,π_0} in the dynamic panel data binary choice model, for different values of ϵ. $n = 500$, results for 1000 simulations. The nominal level for confidence intervals (CI) is 95%.

Table S5: Effect of the PROGRESA subsidy and counterfactual reforms

	Model-based		Minimum-MSE		Experimental	
	PROGRESA impacts					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.076	.080	.077	.054	.087	.050
non-robust CI	(.017,.135)	(.039,.121)	-	-	-	-
robust CI	(-.018,.170)	(-.017,.177)	(.001,.152)	(-.012,.119)	-	-
	Counterfactual 1: doubling subsidy					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.145	.146	.146	.104	-	-
robust CI	(-.022,.312)	(-.011,.304)	(.013,.279)	(.002,.211)	-	-
	Counterfactual 2: unconditional transfer					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.004	.005	.004	-.017	-	-
robust CI	(-.354,.362)	(-.293,.303)	(-.214,.223)	(-.203,.169)	-	-

Notes: Sample from Todd and Wolpin (2006). $\epsilon = \epsilon_1$. CI are 95% confidence intervals. The unconditional transfer amounts to 5000 pesos in a year.

Figure S1: Effect of the PROGRESA subsidy as a function of neighborhood size ϵ



Notes: Sample from Todd and Wolpin (2006). ϵ *is reported 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 model-based prediction. The vertical line indicates* ϵ_1*. Girls (left) and boys (right).*