Minimizing Sensitivity to Model Misspecification*

Stéphane Bonhomme† Martin Weidner‡

October 9, 2018

Abstract

We propose a framework for estimation and inference about the parameters of an economic model and predictions based on it, 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 derive formulas to construct estimators whose mean squared error is minimax in a neighborhood of the reference model, based on simple one-step adjustments. We construct confidence intervals that contain the true parameter under both correct specification and local misspecification. We calibrate the degree of misspecification using a model detection error approach. Our approach allows us to perform systematic sensitivity analysis when the parameter of interest may be partially or irregularly identified. To illustrate our approach we study panel data models where the distribution of individual effects may be misspecified and the number of time periods is small, and we revisit the structural evaluation of a conditional cash transfer program in Mexico.

JEL CODES: C13, C23.
KEYWORDS: Model misspecification, robustness, sensitivity analysis, structural models, counterfactuals, latent variables, panel data.

*We thank Josh Angrist, Tim Armstrong, Gary Chamberlain, Tim Christensen, Ben Connault, 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 the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and from the European Research Council grant ERC-2014-CoG-646917-ROMIA.

†University of Chicago.
‡University College London.

1 Introduction

Although economic models are intended as plausible approximations to a complex economic reality, econometric inference typically relies on the model being an exact description of the population environment. This tension is most salient in the use of structural models to predict the effects of counterfactual policies. Given estimates of model parameters, it is common practice to simply "plug in" those parameters to compute the effect of interest. Such a practice, which typically requires full specification of the economic environment, hinges on the model being correctly specified.

Economists have long recognized the risk of model misspecification. A number of approaches have been developed, such as specification tests and estimation of more general nesting models, semi-parametric and nonparametric methods, and more recently bounds approaches. Implementing those existing 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 which account for model misspecification error in addition to sampling uncertainty.

Our approach is based on considering deviations from a *reference specification* of the model, which 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. While it is theoretically possible to extend our approach to allow for non-local deviations, a local analysis presents important advantages in terms of tractability since it allows us to rely on linearization techniques.

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 borrows several key elements from Hansen and Sargent's (2001, 2008) work on robust decision making under uncertainty and ambiguity. In particular, we rely on their approach to calibrate the size of the neighborhood around the reference model in a way that targets the probability of a model detection error. Our approach thus delivers a class of estimators indexed by error probabilities, which 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. In our approach, 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 mean squared error as well as for the minimum-mean squared error estimators in a given neighborhood of the reference model. A minimum-mean squared error estimator generically takes the form of a one-step adjustment of the prediction 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. In addition we show that, beyond likelihood settings, our approach can be applied to models defined by moment restrictions.

To illustrate our approach we first analyze a linear regression model where the researcher postulates that covariates are exogenous, while contemplating the possibility that this assumption might be violated. The goal is to estimate a regression parameter. The researcher has a set of instruments, which she believes to be valid, but the rank condition may fail to hold. In this case the minimum-MSE estimator interpolates, in a nonlinear fashion, be-

tween the ordinary least squares (OLS) and instrumental variable (IV) estimators. When the first-stage rank condition holds, letting the neighborhood size tend to infinity gives the IV estimator. However, since the minimax rule induces a particular form of regularization of the first-stage matrix (akin to Ridge regression), the minimum-MSE estimator is always well-defined irrespective of the rank condition.

We then apply our approach to two main illustrations. First, we consider a class of panel data models which covers both static and dynamic settings. Our main focus is on average effects, which depend on the distribution of individual effects. The risk of misspecification of this distribution and its dependence on covariates and initial conditions has been emphasized in the literature (e.g., Heckman, 1981). This setting is also of interest since it has been shown that, in discrete choice panel data models, common parameters and average effects often fail to be point-identified (Chamberlain, 2010, Honoré and Tamer, 2006, Chernozhukov *et al.*, 2013), motivating the use of a sensitivity analysis approach. While existing work provides consistency results based on large-n, T asymptotic arguments (e.g., Arellano and Bonhomme, 2009), here we focus on assessing sensitivity to misspecification in a fixed-T setting.

In panel data models, we show that minimizing mean squared error leads to a regularization approach (specifically, Tikhonov regularization). The penalization reflects the degree of misspecification allowed for, which is itself calibrated based on a detection error probability. When the parameter of interest is point-identified and root-n consistently estimable the estimator converges to a semi-parametrically consistent estimator as the neighborhood size tends to infinity. Importantly, our approach remains informative when identification is irregular or point-identification fails. In simulations of a dynamic panel probit model under misspecification, we illustrate that our estimator can provide substantial bias and MSE reduction relative to commonly used estimators.

As a second illustration 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 out. They compare the predictions of the structural model with the estimated experimental impact. As emphasized by Todd and Wolpin (2008) and Attanasio *et al.* (2012), the ability to predict the effects of the program based solely on control villages imposes restrictions on the economic model. Within a simple static model of education choice,

we assess the sensitivity of 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, in which case control villages are no longer sufficient to predict program impacts (Wolpin, 2013). We also provide improved 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.

Related literature. This paper relates to several branches of the literature in econometrics and statistics on robustness and sensitivity analysis. As in the literature on robust statistics dating back to Huber (1964), we rely on a minimax approach and aim to minimize the worst-case impact of misspecification in a neighborhood of a model. See Huber and Ronchetti (2009) for a comprehensive account of this literature. Our approach is closest to the infinitesimal approach based on influence functions (Hampel *et al.*, 1986), and especially to the shrinking neighborhood approach developed by Rieder (1994). An important difference with this previous work, and with recent papers on sensitivity analysis that we mention below, is that we focus on misspecification of *specific aspects* of a model. That is, we consider 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 branch of the literature is the work on orthogonalization and locally robust moment functions, as developed in Neyman (1959), Newey (1994), Chernozhukov *et al.* (2016), and Chernozhukov *et al.* (2018), among others. Similarly to those approaches, we wish to construct estimators which are relatively insensitive to variation in an input. A difference is that we account for both bias and variance, weighting them by calibrating the size of the neighborhood around the reference model. In addition, our approach to robustness and sensitivity – both for estimation and construction of confidence intervals – does not require the larger model to be point-identified. A precedent of the idea of minimum sensitivity is the concept of local unbiasedness proposed by Fraser (1964).

Our analysis is also connected to Bayesian robustness, see for example Berger and Berliner (1986), Gustafson (2000), Vidakovic (2000), or recently Mueller (2012). In our approach we similarly focus on sensitivity to model (or "prior") assumptions. However, our minimum-mean squared error estimators and confidence intervals have a frequentist interpretation.

Closely related to our work is the literature on statistical decision theory dating back to Wald (1950); see for example Chamberlain (2000), Watson and Holmes (2016), and Hansen and Marinacci (2016). Hansen and Sargent (2008) provide compelling motivation for the use of a minimax approach based on Kullback-Leibler neighborhoods whose widths are calibrated based on detection error probabilities.

This paper also relates to the literature on sensitivity analysis in statistics and economics, for example Rosenbaum and Rubin (1983a), Leamer (1985), Imbens (2003), Altonji *et al.* (2005), Nevo and Rosen (2012), Oster (2014), and 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, which relies on a local asymptotic to guide model choice.

Recent papers rely on a local approach to misspecification related to ours to provide tools for sensitivity analysis. Andrews *et al.* (2017) propose a measure of sensitivity of parameter estimates in structural economic models to the moments used in estimation. Andrews *et al.* (2018) introduce a measure of informativeness of descriptive statistics and other reduced-form moments in the estimation of structural models; see also recent work by Mukhin (2018). Our goal is different, in that we aim to provide a framework for estimation and inference in the presence of misspecification. In independent work, 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. In this setting they derive results on optimal estimation and inference. Differently from their approach, here we seek to ensure robustness to misspecification of a reference model (for example, a panel data model with a parametrically specified distribution of individual effects) within a larger class of models (e.g., models with an unrestricted distribution of individual effects).

Our focus on *specific* forms of model misspecification is close in spirit to some recently proposed approaches to estimate partially identified models. Chen *et al.* (2011) and Norets and Tang (2014) develop methods for sensitivity analysis based on estimating semi-parametric models while allowing for non-point identification in inference. Schennach (2013) proposes a related approach in the context of latent variables models. In recent independent work,

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. We view our approach as complementary to these partial identification methods. Our local approach allows tractability in complex models, such as structural economic models, since implementation does not require estimating a larger model. In our framework, parametric reference models are still seen as useful benchmarks, although their predictions need to be modified in order to minimize the impact of misspecification. This aspect relates our paper to shrinkage methods, such as those recently proposed by Hansen (2016, 2017) and Fessler and Kasy (2018); see Maasoumi (1978) for an early contribution. Our approach differs from the shrinkage literature since, instead of estimating an unrestricted estimator and shrinking it towards a set of restrictions, we adjust – in one step – a restricted estimator. Moreover, we calibrate the size of the neighborhood, hence the degree of "shrinkage", rather than attempting to estimate it.

The plan of the paper is as follows. In Section 2 we describe our framework and derive the main results. In Sections 3 and 4 we apply our framework to parametric and semi-parametric likelihood settings, respectively. In Sections 5 and 6 we show the results of a simulation exercise in a panel data model, and the empirical illustration on conditional cash transfers in Mexico. We discuss several extensions in Section 7, and we conclude in Section 8. Three appendices numbered A, B and C provide the proofs, and details on various extensions.

2 Framework of analysis

In this section we describe the main elements of our approach in a general setting. In the next two sections we will specialize the analysis to the cases of parametric misspecification, and semi-parametric misspecification of distributional functional forms.

2.1 Setup

We observe a random sample $(Y_i : i = 1, \ldots, n)$ from the distribution $f_\theta(y) = f(y \,|\, \theta)$, where $\theta \in \Theta$ is a finite- or infinite-dimensional parameter. Throughout the paper the parameter of interest is δ_θ, a scalar function or functional of θ. We assume that δ_θ and f_θ are known, smooth functions of θ. Examples of functionals of interest in economic applications include counterfactual policy effects which can be computed given a fully specified structural model,

and moments of observed and latent data such as average effects in panel data settings. The true parameter value $\theta_0 \in \Theta$ that generates the observed data $Y_1, \ldots, Y_n$ is unknown to the researcher. Our goal is to estimate δ_{θ_0} and construct confidence intervals around it.

Our starting point is that the unknown true θ_0 belongs to a neighborhood of a reference model $\theta(\eta)$, indexed by a finite-dimensional parameter vector $\eta \in \mathcal{B}$. We say that the reference model is *correctly specified* if there is an $\eta \in \mathcal{B}$ such that $\theta_0 = \theta(\eta)$. Otherwise we say that the model is *misspecified*. Note that this setup covers the estimation of (structural) parameters of the reference model as a special case, when η is a component of θ and $\delta_\theta = \eta$.

To quantify the degree of misspecification we rely on a distance measure d on Θ. Let $\mathbb{E}_\theta$ be the expectation under the distribution $\prod_{i=1}^n f_\theta(Y_i)$. We will measure the performance of an estimator $\widehat{\delta}$ by its worst-case bias $|\mathbb{E}_{\theta_0}\widehat{\delta} - \delta_{\theta_0}|$ and mean squared error (MSE) $\mathbb{E}_{\theta_0}[(\widehat{\delta} - \delta_{\theta_0})^2]$ in an *ϵ-neighborhood* Γ_ϵ of the reference model manifold, which is defined as

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

At the end of this section we will discuss how to choose $\epsilon \geq 0$ through a calibration approach.

Examples As a first example, consider a parametric model defined by an Euclidean parameter $\theta \in \Theta$. Under the reference model, θ satisfies a set of restrictions. To fix ideas, let $\theta = (\beta, \rho)$, $\eta = \beta$, and consider the reference specification $\theta(\eta) = (\beta, 0)$, which corresponds to imposing the restriction that $\rho = 0$. For example, ρ 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 3.2. Suppose that the researcher is interested in the parameter $\delta_\theta = c'\beta$ for a known vector c, such as one component of β. In this case we define the neighborhood Γ_ϵ using the weighted Euclidean (squared) distance $d(\theta_0, \theta) = \|\beta_0 - \beta\|^2_{\Omega_\beta} + \|\rho_0 - \rho\|^2_{\Omega_\rho}$, for two positive-definite matrices Ω_β and Ω_ρ, where $\|V\|^2_\Omega = V'\Omega V$. We further analyze this class of models in Section 3.

As a second example, consider a semi-parametric panel data model whose likelihood depends on a finite-dimensional parameter vector β and a nonparametric density π of individual effects $A \in \mathcal{A}$ (abstracting from conditioning covariates for simplicity). The joint density of (Y, A) is $g_\beta(y \,|\, a)\pi(a)$ for some known function g. Suppose that the researcher's goal is to estimate an average effect $\delta_\theta = \mathbb{E}_\pi \Delta(A, \beta)$, for Δ a known function. It is common to estimate the model by parameterizing the unknown density using a correlated random-effects specification π_γ, where γ is finite-dimensional (e.g., a Gaussian whose mean and variance are

the components of γ). We focus on situations where, although the researcher thinks of π_γ 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 let $d(\theta_0, \theta) = \|\beta_0 - \beta\|^2_{\Omega_\beta} + 2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right)\pi_0(a)da$, for a positive-definite matrix Ω_β. We analyze this class of models in Section 4.

We study a local asymptotic framework where ϵ tends to zero and the sample size n tends to infinity. Specifically, we will choose ϵ such that ϵn is asymptotically constant. The reason for focusing on ϵ tending to zero is tractability. While fixed-ϵ minimax calculations involve considerable mathematical difficulties, a small-ϵ analysis allows us to rely on linearization techniques and obtain simple, explicit expressions. Moreover, in an asymptotic where ϵn tends to a constant both bias and variance play a non-trivial role. This approach has a number of precedents in the literature (notably Rieder, 1994).

We will focus on *asymptotically linear* estimators, which can be expanded around $\delta_{\theta(\eta)}$ for a suitable η; that is, for small ϵ and large n the estimators we consider will satisfy

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

where $h(y, \eta) = \phi(y, \theta(\eta))$, for $\phi(y, \theta_0)$ the influence function of $\widehat{\delta}$. We will assume that the remainder in (1) is uniformly small on Γ_ϵ in a sense to be made precise in Theorem 1 below.

In addition, we assume that the function h in (1) satisfies two key conditions. First, it has zero mean under the reference model; that is,

$$\mathbb{E}_{\theta(\eta)} h(Y, \eta) = 0, \quad \text{for all } \eta \in \mathcal{B}, \tag{2}$$

where we write Y to denote Y_i for one representative $i \in \{1, \ldots, n\}$. Under (2), the estimator $\widehat{\delta}$ is asymptotically unbiased for the target parameter $\delta_{\theta_0} = \delta_{\theta(\eta)}$ under the reference model. Second, h is *locally robust* with respect to η in the following sense,

$$\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{\theta(\eta)} \nabla_\eta h(Y, \eta) = 0, \quad \text{for all } \eta \in \mathcal{B}, \tag{3}$$

where ∇_η is the gradient operator. The constraint (3) guarantees that the estimator $\widehat{\delta} = \widehat{\delta}(Y_1, \ldots, Y_n)$ itself does not have an explicit η-dependence, but only depends on the model parameters through the distribution of the sample. By differentiating (2) with respect to η we obtain the following equivalent expression for (3),

$$\mathbb{E}_{\theta(\eta)}\, h(Y, \eta)\, \nabla_\eta \log f_{\theta(\eta)}(Y) = \nabla_\eta \delta_{\theta(\eta)}, \quad \text{for all } \eta \in \mathcal{B}. \tag{4}$$

Local robustness (3)-(4) follows from properties of influence functions under general conditions; see Chernozhukov *et al.* (2016), for example.

Estimators based on moment restrictions or score equations which are satisfied under the reference model (but may not hold under f_{θ_0}) can under mild conditions be expanded as in (1) for a suitable h function satisfying (2) and (3)-(4). In Appendix A we provide more details about the asymptotically linear representation (1), and we give several examples of estimators.[1]

In this paper we characterize the worst-case asymptotic bias and MSE of estimators that satisfy the above conditions, and construct confidence intervals for the target parameter δ_{θ_0} which are uniformly asymptotically valid on the neighborhood Γ_ϵ. In addition, an important goal of the analysis is to construct estimators that are asymptotically optimal in a minimax sense. For this purpose, we will show how to compute a function h such that the worst-case MSE, in the neighborhood Γ_ϵ, among estimators of the form

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

is minimized under our local asymptotic analysis. Here $\widehat{\eta}$ is a preliminary estimator of η, for example the maximum likelihood estimator (MLE) of η based on the reference model. In fact, it follows from the local robustness property (3) that, under mild conditions on the preliminary estimator, $\widehat{\delta}_{h,\widehat{\eta}}$ satisfies (1) for that same function h. As a result, the form of the minimum-MSE h function will not be affected by the choice of $\widehat{\eta}$.

Examples (cont.) In our first, parametric example a natural estimator is the MLE of $c'\beta$ based on the reference specification, for example, the OLS estimator under the assumption that ρ – 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 second, semi-parametric example a commonly used ("random-effects") estimator of $\delta_\theta = \mathbb{E}_\pi \Delta(A, \beta)$ 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 ("em-

[1]Note that, in (1), the estimator is expanded around the reference value $\delta_{\theta(\eta)}$. As we discuss in Appendix A, such asymptotic expansions can be related to expansions around the probability limit of $\widehat{\delta}$ under f_{θ_0} – i.e., around the "pseudo-true value" of the target parameter.

pirical Bayes") estimator is obtained by substituting an integral with respect to the posterior distribution of individual effects based on $\pi_{\widehat{\gamma}}$. 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, whenever π_0 does not belong to the postulated parametric family π_γ. We compare their finite-sample performance to that of our minimum-MSE estimator in Section 5.

2.2 Heuristic derivation of the minimum-MSE estimator

We start by providing heuristic derivations of worst-case bias and minimum-MSE estimator. This will lead to the main definitions in equations (8), (11) and (12) below. Then, in the next subsection, we will provide regularity conditions under which these derivations are formally justified.

For presentation purposes we first describe our approach in the simple case where the parameter η, and hence the reference model $\theta(\eta)$, are known; that is, we assume that $\mathcal{B} = \{\eta\}$. For any $\epsilon \geq 0$, let

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

We assume that Θ and $\Gamma_\epsilon(\eta)$ are convex sets. For any linear map $u : \Theta \to \mathbb{R}$ we define

$$\|u\|_{\eta,\epsilon} = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \epsilon^{-\frac{1}{2}}\, u'(\theta_0 - \theta(\eta)), \quad \|u\|_\eta = \lim_{\epsilon \to 0} \|u\|_{\eta,\epsilon}\,. \tag{6}$$

When θ is infinite-dimensional this definition continues to hold, with a suitable ("bracket") notation for $u'(\theta_0 - \theta(\eta))$; see Appendix A for a general notation that covers both finite and infinite-dimensional cases. We assume that the distance measure d is chosen such that $\|\cdot\|_\eta$ is unique and well-defined, and that it constitutes a norm. $\|\cdot\|_\eta$ is *dual* to a local approximation of $d(\theta_0, \theta(\eta))$ for fixed $\theta(\eta)$. 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) holds. Under appropriate regularity conditions, the worst-case bias of $\widehat{\delta}$ in the neighborhood $\Gamma_\epsilon(\eta)$ can be expanded for small ϵ and large n as

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \left| \mathbb{E}_{\theta_0} \widehat{\delta} - \delta_{\theta_0} \right| = b_\epsilon(h, \eta) + o(\epsilon^{\frac{1}{2}}) + o(n^{-\frac{1}{2}}), \tag{7}$$

where

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

for $\|\cdot\|_\eta$ the dual norm defined in (6). When θ is infinite-dimensional ∇_θ denotes a general (Gâteaux) derivative . Then, the worst-case MSE in $\Gamma_\epsilon(\eta)$ can be expanded as follows, again under appropriate regularity conditions,

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta} - \delta_{\theta_0}\right)^2\right] = b_\epsilon(h,\eta)^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n} + o(\epsilon) + o\left(n^{-1}\right). \tag{9}$$

In order to construct estimators with minimum worst-case MSE we define, for any function h satisfying (2),

$$\widehat{\delta}_{h,\eta} = \delta_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i,\eta). \tag{10}$$

Applying the small-ϵ approximation of the bias and MSE to $\widehat{\delta}_{h,\eta}$, we define the *minimum-MSE* function $h_\epsilon^{\mathrm{MMSE}}(y,\eta)$ as

$$h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta) = \underset{h(\cdot,\eta)}{\mathrm{argmin}}\ \epsilon \left\|\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)}\, h(Y,\eta)\ \nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n}$$
$$\text{subject to (2)}. \tag{11}$$

The minimum-MSE estimator $\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \delta_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta)$ thus minimizes an asymptotic approximation to the worst-case MSE in $\Gamma_\epsilon(\eta)$. Using a small-ϵ 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 (11) is convex.

Note that, for $\epsilon = 0$ we have $\widehat{\delta}_0^{\mathrm{MMSE}} = \delta_{\theta(\eta)}$, independent of the data, since this choice satisfies the unbiasedness constraint and achieves zero variance. However, for $\epsilon > 0$ the minimum-MSE function $h_\epsilon^{\mathrm{MMSE}}(y,\eta)$ depends on y, hence the estimator $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ depends on the data $Y_1,\ldots,Y_n$.[2]

Turning now to the general case where the parameter η is unknown, let $\widehat{\eta}$ be a preliminary estimator of η that is asymptotically unbiased for η under the reference model $f_{\theta(\eta)}$. Let $h(\cdot,\eta)$ be a set of functions indexed by η, and define $\widehat{\delta}_{h,\widehat{\eta}}$ by (5). We assume that, in addition to (2), $h(\cdot,\eta)$ satisfies the local robustness condition (4). Analogously to (11), we search for functions $h(\cdot,\eta)$ solving the following programs, separately for all $\eta \in \mathcal{B}$,

$$h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta) = \underset{h(\cdot,\eta)}{\mathrm{argmin}}\ \epsilon \left\|\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)}\, h(Y,\eta)\ \nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n}$$
$$\text{subject to (2) and (4)}, \tag{12}$$

[2]The function $h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta)$ also depends on the sample size n, although we do not make the dependence explicit. In fact, $h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta)$ only depends on ϵ and n through the product ϵn.

where we note that (12) is again a convex optimization problem.

We then define the minimum-MSE estimator of δ_{θ_0} as

$$\widehat{\delta}_{\epsilon}^{\text{MMSE}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^{n} h_{\epsilon}^{\text{MMSE}}(Y_i, \widehat{\eta}). \tag{13}$$

In practice, (12) only needs to be solved at $\eta = \widehat{\eta}$. In addition, the form of the minimum-MSE estimator is not affected by the choice of the preliminary estimator $\widehat{\eta}$.

It is common in applications with covariates to model the conditional distributions of outcomes Y given covariates X as $f_\theta(y \,|\, x)$, while leaving the marginal distribution of X, $f_X(x)$, unspecified. Our approach can easily be adapted to deal with such conditional models. In those cases we minimize the (worst-case) conditional MSE

$$\mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_{h,\widehat{\eta}} - \delta_{\theta_0}\right)^2 \,\middle|\, X_1, \ldots, X_n\right],$$

for estimators $\widehat{\delta}_{h,\widehat{\eta}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i, X_i, \widehat{\eta})$. The calculations for $h_{\epsilon}^{\text{MMSE}}$ and $\widehat{\delta}_{\epsilon}^{\text{MMSE}}$ are very similar in this case, as we will see in the parametric and semi-parametric settings of Sections 3 and 4.

Special cases. To provide intuition on the minimum-MSE function $h_{\epsilon}^{\text{MMSE}}$, let us define two Hessian matrices $H_{\theta(\eta)}$ ($\dim\theta \times \dim\theta$) and H_η ($\dim\eta \times \dim\eta$) as

$$H_{\theta(\eta)} = \mathbb{E}_{\theta(\eta)}\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]', \; H_\eta = \mathbb{E}_{\theta(\eta)}\left[\nabla_\eta \log f_{\theta(\eta)}(Y)\right]\left[\nabla_\eta \log f_{\theta(\eta)}(Y)\right]'.$$

The definition of $H_{\theta(\eta)}$ generalizes to the infinite-dimensional θ case, see Appendix A.

In our analysis we assume that H_η is invertible. This requires that the Hessian matrix of the parametric reference model be non-singular, thus requiring that η be identified under the reference model. For $\epsilon = 0$ we find that

$$h_0^{\text{MMSE}}(y, \eta) = \left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)}. \tag{14}$$

Thus, if we impose that $\epsilon = 0$ – that is, if we work under the assumption that the parametric reference model is correctly specified – then $\widehat{\delta}_{\epsilon}^{\text{MMSE}}$ is simply the one-step approximation of the MLE for δ_{θ_0} that maximizes the likelihood with respect to the "small" parameter η. This "one-step efficient" adjustment of $\delta_{\theta(\widehat{\eta})}$ is purely based on efficiency considerations.[3]

[3]Such one-step approximations are classical estimators in statistics; see for example Bickel *et al.* (1993, pp. 43–45).

Another interesting special case of the minimum-MSE h function arises in the limit $\epsilon \to \infty$, when the matrix or operator $H_{\theta(\eta)}$ is invertible. Note that invertibility of $H_{\theta(\eta)}$, which may fail when θ_0 is not identified, is not needed in our analysis and we only use it to analyze this special case. We then have that

$$\lim_{\epsilon \to \infty} h_\epsilon^{\text{MMSE}}(y, \eta) = \left[\nabla_\theta \log f_{\theta(\eta)}(y)\right]' \, H_{\theta(\eta)}^{-1} \, \nabla_\theta \delta_{\theta(\eta)}. \tag{15}$$

Equivalently, the same limiting quantity is attained if ϵ is kept fixed as $n \to \infty$, or if ϵn tends to infinity. In this limit we thus find that $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is simply the one-step approximation of the MLE for δ_{θ_0} that maximizes the likelihood with respect to the "large" parameter θ.

More generally, 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 (12) is only a *local* approximation to the full MSE-minimization problem. When $H_{\theta(\eta)}$ is invertible it can be shown that $b_\epsilon(h_\epsilon^{\text{MMSE}}(\cdot, \eta), \eta)$ tends to zero as ϵ tends to infinity, since the one-step MLE approximation of the large model is robust to misspecification of $f_{\theta(\eta)}$. Lastly, note that, while neither (14) nor (15) 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.

The estimator associated with (15) is "orthogonalized" or "locally robust" (e.g., Neyman, 1959, Chernozhukov *et al.*, 2016) with respect to the large parameter θ.[4] While such estimators are useful in a number of settings, in our framework they have minimal bias but may have large variance. As a result they may be ill-behaved in non point-identified problems, or in problems where the identification of θ_0 is irregular. By contrast, notice that when $H_{\theta(\eta)}$ is singular $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is still well-defined and unique, due to the variance of $h(Y, \eta)$ acting as a sample size-dependent regularization. The form of $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is thus based on both efficiency and robustness considerations.

Examples (cont.). To describe the form of the bias and MSE in our two examples, consider first a parametric model with distance measure $d(\theta_0, \theta) = \|\theta_0 - \theta\|_\Omega^2$. Any linear map on Θ can be written as the transpose of a $\dim \theta$-dimensional vector u, and we have

$$\|u\|_{\eta,\epsilon} = \|u\|_\eta = \|u\|_{\Omega^{-1}},$$

[4] To see this, it is useful to explicitly indicate the dependence of h on θ. The moment condition $\mathbb{E}_\theta(\delta_\theta + h(Y, \theta) - \delta) = 0$ is locally robust with respect to θ whenever $\mathbb{E}_\theta \nabla_\theta(\delta_\theta + h(Y, \theta)) = 0$. The function $h(y, \theta) = [\nabla_\theta \log f_\theta(y)]' \, H_\theta^{-1} \, \nabla_\theta \delta_\theta$ is locally robust in this sense.

where Ω^{-1} is the inverse of Ω. The squared bias term in (12) is then a quadratic function of h, and computing $h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta)$ amounts to minimizing a quadratic objective in h. In Section 3 we will see that this problem has a closed-form solution.

Consider next our semi-parametric example, abstracting from β parameters and taking $\theta = \pi$ for simplicity, with distance measure $d(\theta_0, \theta) = 2\int_{\mathcal{A}} \log\left(\frac{\theta_0(a)}{\theta(a)}\right)\theta_0(a)da$. We show in Appendix B that for any real-valued function $q : \mathcal{A} \to \mathbb{R}$ associated with the linear map $\theta \mapsto \int_{\mathcal{A}} q(a)\theta(a)da$ we have, under mild conditions,

$$\|q\|_\eta = \sqrt{\mathrm{Var}_{\theta(\eta)}\left(q(A)\right)}. \tag{16}$$

Moreover, in settings where f_θ and δ_θ are linear in θ, the derivatives $\nabla_\theta \delta_{\theta(\eta)}$ and $\nabla_\theta \log f_{\theta(\eta)}(y)$ take the form of simple, analytical expressions. Indeed, using that $\delta_\theta = \mathbb{E}_\theta \Delta(A)$, $f_\theta(y) = \int_{\mathcal{A}} g(y \,|\, a)\theta(a)da$, and $\int_{\mathcal{A}} \theta(a)da = 1$, we have (see Appendix B for a formal presentation)

$$\nabla_\theta \delta_\theta = \Delta(\cdot) - \delta_\theta, \quad \nabla_\theta \log f_\theta(y) = \frac{g(y \,|\, \cdot)}{\int_{\mathcal{A}} g(y \,|\, a)\theta(a)da} - 1.$$

It thus follows that, for h satisfying (2),

$$\mathbb{E}_{\theta(\eta)}\, h(Y, \eta)\, \nabla_\theta \log f_{\theta(\eta)}(Y) = \int_{\mathcal{Y}} h(y, \eta) g(y \,|\, \cdot)dy = \mathbb{E}\left[h(Y, \eta) \,|\, A = \cdot\right].$$

For example, (8) and (11) become, respectively,

$$b_\epsilon(h, \eta) = \epsilon^{\frac{1}{2}} \sqrt{\mathrm{Var}_{\theta(\eta)}\left(\Delta(A) - \mathbb{E}\left[h(Y, \eta) \,|\, A\right]\right)}, \tag{17}$$

and

$$h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta) = \underset{h(\cdot,\eta)}{\mathrm{argmin}}\ \epsilon\, \mathrm{Var}_{\theta(\eta)}\left(\Delta(A) - \mathbb{E}\left[h(Y, \eta) \,|\, A\right]\right) + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y, \eta))}{n}$$
$$\text{subject to (2)}. \tag{18}$$

As in the parametric case, the MSE-minimization problem (18) is thus quadratic in h, and computing $h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta)$ amounts to solving a quadratic problem.

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 large class of estimators as n tends to infinity and ϵn tends to a constant. Moreover, under the stated assumptions the heuristic derivations of the previous subsection are formally justified.

We will show that the minimum-MSE estimator asymptotically minimizes the following integrated worst-case MSE,

$$\int_{\mathcal{B}} \left\{ \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_{h,\widehat{\eta}} - \delta_{\theta_0} \right)^2 \right] \right\} w(\eta) d\eta, \tag{19}$$

where w is a non-negative weight function supported on $\mathcal{B}$. This particular objective has the advantage, compared to minimizing the maximum MSE on the set of (θ_0, η) in Γ_ϵ, of not being driven by the worst-case MSE in terms of η values. Moreover, the optimization problem in (19) nicely decouples across η asymptotically, and its solution does not depend on the weight function w.

We first establish the following result. All proofs are in Appendix A.

Theorem 1. *Let Assumptions A1 and A2 in Appendix A hold, and let* $\widehat{\delta}_\epsilon = \widehat{\delta}_\epsilon(Y_1, \ldots, Y_n)$ *be a sequence of estimators such that*

$$\sup_{(\theta_0,\eta) \in \Gamma_\epsilon} \mathbb{E}_{\theta_0} \left[\widehat{\delta}_\epsilon - \delta_{\theta(\eta)} - \frac{1}{n} \sum_{i=1}^{n} h_\epsilon(Y_i, \eta) \right]^2 = o(\epsilon), \tag{20}$$

for a sequence of influence functions $h_\epsilon(\cdot, \eta)$ *that satisfy the constraints (2) and (4), as well as* $\sup_{(\theta_0,\eta) \in \Gamma_\epsilon} \mathbb{E}_{\theta_0} |h_\epsilon(Y, \eta)|^\kappa = O(1)$*, for some* $\kappa > 2$*. We then have*

$$\sup_{\eta \in \mathcal{B}} \left\{ \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon^{\,\mathrm{MMSE}} - \delta_{\theta_0} \right)^2 \right] - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0} \right)^2 \right] \right\} \leq o(\epsilon). \tag{21}$$

Theorem 1 is established in a joint asymptotic where ϵ tends to zero as n tends to infinity and ϵn tends to a finite positive constant. The sequences of estimators and influence functions could thus alternatively be indexed by n. 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 states that the leading order worst-case MSE achieved by our minimum-MSE estimator $\widehat{\delta}_\epsilon^{\,\mathrm{MMSE}}$ is at least as good as the leading order worst-case MSE achieved by any other sequence of estimators satisfying our regularity conditions. All the assumptions on $\widehat{\delta}_\epsilon$ and $h_\epsilon(y, \eta)$ that we require for this result are explicitly listed in the statement of the theorem. In particular, condition (20) is a form of local regularity of the sequence of estimators $\widehat{\delta}_\epsilon$ (e.g., Bickel *et al.*, 1993). The additional regularity conditions in Assumptions A1 and A2 are smoothness conditions on $f_{\theta_0}(y)$, δ_{θ_0}, $\theta(\eta)$, and $d(\theta_0, \theta(\eta))$ as functions of θ_0 and η, and an appropriate rate condition on the preliminary estimator $\widehat{\eta}$.

The optimality result in Theorem 1 is uniform in the reference parameter η. Such a uniform result is possible here, because our constraints (2) and (4) imply a decoupling of the worst-case MSE optimization problem across η; that is, we can solve for the optimal $h_\epsilon^{\text{MMSE}}(\cdot,\eta)$ separately for each value of η. This happens since (2), (4) and (9) only involve $h(\cdot,\eta)$ at a given η value, and since $\widehat{\delta}_{h,\widehat{\eta}}$ satisfies (1) under local robustness.[5]

To leading order, the uniform optimality result in Theorem 1 immediately implies the following corollary on the integrated worst-case MSE.

Corollary 1. *Under the Assumptions of Theorem 1 we also have*

$$\int_{\mathcal{B}} \left\{ \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\theta_0} \right)^2 \right] \right\} w(\eta) d\eta \leq \int_{\mathcal{B}} \left\{ \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0} \right)^2 \right] \right\} w(\eta) d\eta + o(\epsilon),$$

for any weight function $w : \mathcal{B} \to [0,\infty)$ that satisfies $\int_{\mathcal{B}} w(\eta)d\eta < \infty$.

2.4 Confidence intervals

In addition to point estimates, our framework allows us to compute confidence intervals that contain δ_{θ_0} with prespecified probability under our local asymptotic. To see this, let $\widehat{\delta}$ be an estimator satisfying (1), (2) and (4). For a given confidence level $\mu \in (0,1)$, let us define the following interval

$$CI_\epsilon(1-\mu, \widehat{\delta}) = \left[\widehat{\delta} \pm \left(b_\epsilon\left(h, \widehat{\eta}\right) + \frac{\widehat{\sigma}_h}{\sqrt{n}} c_{1-\mu/2} \right) \right], \tag{22}$$

where $b_\epsilon(h,\eta)$ is given by (8), $\widehat{\sigma}_h^2$ is the sample variance of $h(Y_1,\widehat{\eta}),\ldots,h(Y_n,\widehat{\eta})$, and $c_{1-\mu/2} = \Phi^{-1}(1-\mu/2)$ is the $(1-\mu/2)$-standard normal quantile. Under suitable regularity conditions, the interval $CI_\epsilon(1-\mu,\widehat{\delta})$ contains δ_{θ_0} with probability approaching $1-\mu$ 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 Assumptions A1 and A3 in Appendix A hold, and also assume that the influence function h of $\widehat{\delta}$ satisfies $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} h^2(Y,\eta) = O(1)$. Then we have*

$$\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \Pr_{\theta_0}\left[\delta_{\theta_0} \in CI_\epsilon(1-\mu,\widehat{\delta}) \right] \geq 1-\mu+o(1). \tag{23}$$

[5]This decoupling only occurs for the leading terms of order ϵ and $1/n$ in the worst-case MSE. If we considered higher-order MSE terms, or even a finite-sample problem, then minimizing the integrated worst-case MSE in (19) would not lead to such decoupling.

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).[6]

2.5 Choice of ϵ

Confidence intervals and minimum-MSE estimators depend on the choice of the neighborhood size ϵ. To provide a meaningful interpretation for this choice we follow a similar *calibration approach* as Hansen and Sargent (2008), and target the probability of a model detection error. For $\theta_0 \in \Theta$ and $\eta \in \mathcal{B}$, consider the following probability of detection error

$$e(\theta_0, \theta(\eta)) = \frac{1}{2}\left\{ \Pr_{\theta_0}\left[\sum_{i=1}^{n} \log\left(\frac{f_{\theta(\eta)}(Y_i)}{f_{\theta_0}(Y_i)}\right) > 0\right] + \Pr_{\theta(\eta)}\left[\sum_{i=1}^{n} \log\left(\frac{f_{\theta_0}(Y_i)}{f_{\theta(\eta)}(Y_i)}\right) > 0\right]\right\}.$$

The function $e(\theta_0, \theta(\eta))$, which is symmetric in its arguments, is an average of two error probabilities corresponding to the data being generated under f_{θ_0} or $f_{\theta(\eta)}$.

Let $p \in (0,1)$ be a fixed probability, and let $\eta \in \mathcal{B}$. In the known-η case we set ϵ such that

$$\inf_{\theta_0 \in \Gamma_\epsilon(\eta)} e(\theta_0, \theta(\eta)) = p + o(1). \tag{24}$$

In the estimated-η case we denote $\overline{e}(\theta_0, \theta(\cdot)) = \sup_{\eta \in \mathcal{B}} e(\theta_0, \theta(\eta))$, and we set ϵ such that

$$\inf_{\theta_0 \in \Gamma_\epsilon(\eta)} \overline{e}(\theta_0, \theta(\cdot)) = p + o(1). \tag{25}$$

According to this rule, the probability of detection error when attempting to distinguish any element $\theta_0 \in \Gamma_\epsilon(\eta)$ from the reference model is no smaller than p. Moreover, achieving a lower p requires setting a larger ϵ.

Let $\widehat{\eta}$ be a preliminary estimator of η. Expanding (25) as n tends to infinity, a possible choice for ϵ is obtained by solving

$$\sup_{\theta_0 \in \Gamma_\epsilon(\widehat{\eta})} (\theta_0 - \theta(\widehat{\eta}))' \, \widetilde{H}_{\theta(\widehat{\eta})} \, (\theta_0 - \theta(\widehat{\eta})) = \frac{4\left(\Phi^{-1}(p)\right)^2}{n}, \tag{26}$$

where $\widetilde{H}_{\theta(\eta)} = H_{\theta(\eta)} - H_{\theta(\eta)} G_\eta' H_\eta^{-1} G_\eta H_{\theta(\eta)}$, for $G_\eta = \nabla_\eta \theta(\eta)'$ (which is $\dim\theta \times \dim\eta$). In the known-$\eta$ case we obtain a similar formula, with η in place of $\widehat{\eta}$ and $H_{\theta(\eta)}$ in place of $\widetilde{H}_{\theta(\widehat{\eta})}$. Note that this calibration of ϵ is not based on the sample $Y_1, ..., Y_n$. We will see that the

[6] A variation suggested by these authors, which reduces the length of the interval, is to compute the interval as $\widehat{\delta} \pm b_\epsilon(h, \widehat{\eta})$ times the $(1-\mu)$-quantile of $|\mathcal{N}(1, \frac{\widehat{\sigma}_h^2}{b_\epsilon(h,\widehat{\eta})^2 n})|$.

value of ϵ implied by (26) has a closed-form or easily computable expression as a function of p in the parametric and semi-parametric models we will analyze in the next two sections.

Our goal here is to provide an optimal estimator for a given amount of misspecification, which is itself calibrated to the ability to detect deviations from the reference model. We do not aim to tailor the amount of misspecification to a given estimator. This aspect differs from the original Hansen and Sargent approach, which is based on decision-specific worst cases. While one could adopt such an approach to calibrate ϵ,[7] we prefer to calibrate a single model-specific value that can be used to compare different estimators.

Setting $\epsilon = \epsilon(p)$ according to (26) is motivated by a desire to calibrate the fear of misspecification of the researcher. When p is fixed to 1% or 5%, say, values θ_0 inside the neighborhood $\Gamma_\epsilon(\widehat{\eta})$ are hard to statistically distinguish from the reference model based on a sample of n observations. Moreover, for fixed p the product ϵn tends to a constant asymptotically. This approach 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". Calibrating ϵ based on model detection error, as we do, provides an interpretable metric to assess how "large" or "small" a given deviation is.

Given an estimator $\widehat{\delta}$, our framework delivers a collection of confidence intervals $CI_{\epsilon(p)}(1-\mu, \widehat{\delta})$ for different p levels. Reporting those allows one to conduct a sensitivity analysis for any given estimator to possible misspecification of the reference model. In addition, our approach delivers a collection of minimum-MSE estimators $\widehat{\delta}^{\text{MMSE}}_{\epsilon(p)}$ for different p. In practice, it can be informative to report the full sets of $\widehat{\delta}^{\text{MMSE}}_{\epsilon(p)}$ and associated confidence intervals as a function of p, along with the estimator and interval corresponding to a preferred p level. We will report such quantities in our empirical illustration in Section 6.

[7] In our first (parametric) example, the worst-case θ_0 values in (7) are, up to lower-order terms,

$$\theta_0^*(h,\eta,\epsilon) = \theta(\eta) \pm \epsilon^{\frac{1}{2}} \frac{\Omega^{-1}\left(\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y)\right)}{\|\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y)\|_{\Omega^{-1}}}.$$

This motivates the following estimator-specific calibration

$$\epsilon = \frac{4\Phi^{-1}(p)^2}{n} \frac{\|\nabla_\theta \delta_{\theta(\widehat{\eta})} - \mathbb{E}_{\theta(\widehat{\eta})} h(Y,\widehat{\eta}) \nabla_\theta \log f_{\theta(\widehat{\eta})}(Y)\|^2_{\Omega^{-1}}}{\|\nabla_\theta \delta_{\theta(\widehat{\eta})} - \mathbb{E}_{\theta(\widehat{\eta})} h(Y,\widehat{\eta}) \nabla_\theta \log f_{\theta(\widehat{\eta})}(Y)\|^2_{\Omega^{-1}\widetilde{H}_{\theta(\widehat{\eta})}\Omega^{-1}}}.$$

In this case h^{MMSE}_ϵ and ϵ are jointly determined.

It should be noted that our choice of ϵ is *not* based on *a priori* information on the true parameter value or the bias of a given estimator. Our approach thus differs from sensitivity analysis methods which rely on prior information about the parameter of interest. Even in the absence of such information, a variety of other approaches could be used to calibrate ϵ (see Appendix C for an example). Given an alternative rule for the choice of ϵ under which ϵn tends asymptotically to a constant, all other ingredients of our approach would remain identical.

3 Parametric models

In this section and the next we specialize our framework to two leading classes of applications. Here we study the case where θ is finite-dimensional and the distance measure is based on a weighted Euclidean metric $\|\cdot\|_\Omega$ for a positive definite weight matrix Ω. We start by treating Ω and the neighborhood size ϵ as known, before discussing how to choose them in practice.

3.1 Minimum-MSE estimator

In the case where θ is finite-dimensional and the distance measure is based on $\|\cdot\|_\Omega$, the small-ϵ approximation to the bias of $\widehat{\delta}$ is given by (8), with $\|\cdot\|_\eta = \|\cdot\|_{\Omega^{-1}}$. This expression can be used to construct confidence intervals, as we explained in Subsection 2.4. Moreover, the objective function in (12) is quadratic and its solution satisfies

$$\begin{aligned} h_\epsilon^{\mathrm{MMSE}}(y,\eta) = &\left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} \\ &+ (\epsilon n)\left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \Omega^{-1} \left(\widetilde{\nabla}_\theta \delta_{\theta(\eta)} - \mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(Y)\right]\right), \end{aligned} \tag{27}$$

where $\widetilde{\nabla}_\theta = \nabla_\theta - H_{\theta(\eta)} G_\eta' H_\eta^{-1} \nabla_\eta$ is a projected gradient operator, and we have assumed that H_η – the Hessian with respect to the "small" parameter η – is non-singular.

This minimum-MSE h function can equivalently be written as

$$\begin{aligned} h_\epsilon^{\mathrm{MMSE}}(y,\eta) = &\left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} \\ &+ \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \left[\widetilde{H}_{\theta(\eta)} + (\epsilon n)^{-1}\Omega\right]^{-1} \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \end{aligned} \tag{28}$$

for $\widetilde{H}_{\theta(\eta)} = \mathrm{Var}\left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right] = H_{\theta(\eta)} - H_{\theta(\eta)} G_\eta' H_\eta^{-1} G_\eta H_{\theta(\eta)}$. In addition to the "one-step efficient" adjustment $h_0^{\mathrm{MMSE}}(\cdot,\eta)$ given by (14), the minimum-MSE function $h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta)$

provides a further adjustment that is motivated by robustness concerns. In the special case where η is known the expression becomes

$$h_\epsilon^{\text{MMSE}}(y,\eta) = \left[\nabla_\theta \log f_{\theta(\eta)}(y)\right]' \left[H_{\theta(\eta)} + (\epsilon n)^{-1}\Omega\right]^{-1} \nabla_\theta \delta_{\theta(\eta)}. \tag{29}$$

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 (15),

$$\lim_{\epsilon\to\infty} h_\epsilon^{\text{MMSE}}(y,\eta) = \left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \widetilde{H}^\dagger_{\theta(\eta)} \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \tag{30}$$

where $\widetilde{H}^\dagger_{\theta(\eta)}$ denotes the Moore-Penrose generalized inverse of $\widetilde{H}_{\theta(\eta)}$.[8] Comparing (30) and (28) shows that the optimal $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is a regularized version of the one-step full MLE, where $(\epsilon n)^{-1}\Omega$ regularizes the projected Hessian matrix $\widetilde{H}_{\theta(\eta)}$. Our "robust" adjustment remains well-defined when $H_{\theta(\eta)}$ is singular, and it accounts for small or zero eigenvalues of the Hessian in a way that is optimal in terms of worst-case mean squared error.

Choice of ϵ and Ω. To calibrate ϵ for a given weight matrix Ω, we rely on (26), which here simplifies to

$$\sup_{v\in\mathbb{R}^{\dim\theta}\,:\,v'\Omega v\le\epsilon} v'\widetilde{H}_{\theta(\widehat{\eta})} v = \frac{4\left(\Phi^{-1}(p)\right)^2}{n}, \tag{31}$$

the solution of which is

$$\epsilon(p) = \frac{4\left(\Phi^{-1}(p)\right)^2}{n \cdot \lambda_{max}(\Omega^{-\frac{1}{2}}\widetilde{H}_{\theta(\widehat{\eta})}\Omega^{-\frac{1}{2}})}, \tag{32}$$

where $\lambda_{max}(A)$ is the maximal eigenvalue of matrix A.

Our approach also depends on the choice of Ω. One may provide guidance on this choice using a calibration approach related to the one we use for ϵ. To see this, let us focus on $\Omega = \text{diag}(\omega_1, ..., \omega_{\dim\theta})$ being diagonal. Applying the same formula as in (31), but now only considering the deviations $v = \theta_0 - \theta(\eta)$ along the j-th component θ_j, we obtain $\omega_j = \omega \cdot (\widetilde{H}_{\theta(\widehat{\eta})})_{(j,j)}$, the j-th diagonal element of $\widetilde{H}_{\theta(\widehat{\eta})}$ multiplied by some constant ω (which can be chosen equal to one without loss of generality). This provides a possible scaling for the components of θ.

Incorporating covariates. In models with conditioning covariates whose distribution is unspecified, the minimum-MSE h function takes a similar form to the expressions derived

[8] In fact, $\widetilde{H}^\dagger_{\theta(\eta)}$ in (30) can be replaced by any generalized inverse of $\widetilde{H}_{\theta(\eta)}$.

above, except that it involves averages over the covariates sample $X_1, \ldots, X_n$. For example, when minimizing the worst-case conditional MSE, (28) becomes

$$h_\epsilon^{\text{MMSE}}(y, x, \eta) = \left[\nabla_\eta \log f_{\theta(\eta)}(y \,|\, x)\right]' \left(\widehat{\mathbb{E}}_X H_\eta\right)^{-1} \nabla_\eta \delta_{\theta(\eta)} + \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y \,|\, x)\right]' \left[\widehat{\mathbb{E}}_X \widetilde{H}_{\theta(\eta)} + (\epsilon n)^{-1}\Omega\right]^{-1} \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \quad (33)$$

where $\widehat{\mathbb{E}}_X \widetilde{H}_{\theta(\eta)} = \widehat{\mathbb{E}}_X H_{\theta(\eta)} - \widehat{\mathbb{E}}_X H_{\theta(\eta)} G_\eta' \left(\widehat{\mathbb{E}}_X H_\eta\right)^{-1} G_\eta \widehat{\mathbb{E}}_X H_{\theta(\eta)}$, for

$$\widehat{\mathbb{E}}_X H_{\theta(\eta)} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta(\eta)}\left[\nabla_\theta \log f_{\theta(\eta)}(Y \,|\, X_i)\right]\left[\nabla_\theta \log f_{\theta(\eta)}(Y \,|\, X_i)\right]',$$
$$\widehat{\mathbb{E}}_X H_\eta = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta(\eta)}\left[\nabla_\eta \log f_{\theta(\eta)}(Y \,|\, X_i)\right]\left[\nabla_\eta \log f_{\theta(\eta)}(Y \,|\, X_i)\right]'.$$

3.2 A linear regression example

Although we view our approach to be most useful in structural or semi-structural settings where the researcher relies on a rich and tightly specified model, studying a linear model helps illustrate some of the main features of our approach in a simple, transparent setup. Specifically, here we consider the linear regression model

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

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 Π is a $\dim X \times \dim Z$ matrix. We assume that

$$U = \rho' V + \xi,$$

where ξ is normal with zero mean and variance σ^2, independent of X and Z. Let Σ_V, Σ_Z and Σ_X be the covariance matrices of V, Z and X. We assume that Σ_V is non-singular. For simplicity we assume that Π, Σ_V, Σ_Z and σ^2 are known. The parameters are thus $\theta = (\beta, \rho)$. As a reference model we take $\eta = \beta$ and $\theta(\eta) = (\beta, 0)$. That is, the reference model treats X as exogenous, while the larger model allows for endogeneity. The target parameter is $\delta_\theta = c'\beta$ for a known $\dim \beta \times 1$ vector c. Lastly, as a weight matrix Ω we take a block-diagonal matrix with β-block Ω_β and ρ-block Ω_ρ.

From (28) we have[9]

$$h_\epsilon^{\text{MMSE}}(y,x,z,\beta) = (y - x'\beta)x'\Sigma_X^{-1}\, c \\ - (y - x'\beta)\left[(x - \Pi z) - \Sigma_V\Sigma_X^{-1}x\right]'\left[\Sigma_V - \Sigma_V\Sigma_X^{-1}\Sigma_V + (\epsilon n)^{-1}\Omega_\rho\right]^{-1}\Sigma_V\Sigma_X^{-1}c. \tag{34}$$

Hence, when $\epsilon = 0$ the minimum-MSE estimator of $c'\beta$ is the "one-step efficient" adjustment in the direction of the OLS estimator, with h function

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

As ϵ tends to infinity, provided $\Pi\Sigma_Z\Pi'$ is invertible, the adjustment is performed in the direction of the IV estimator.[10] Indeed, it follows from (34) that

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

For given $\epsilon > 0$ and n, our adjustment remains well-defined even when $\Pi\Sigma_Z\Pi'$ is singular. When $c'\beta$ is identified (that is, when c belongs to the range of Π) 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 (34) acts as a form of regularization, akin to Ridge regression.

Lastly, for a probability p of model detection error, the choice of ϵ is given by (32); that is,

$$\epsilon(p) = \frac{4\sigma^2\left(\Phi^{-1}(p)\right)^2}{n\cdot\lambda_{max}\left(\Omega_\rho^{-\frac{1}{2}}\left(\Sigma_V - \Sigma_V\Sigma_X^{-1}\Sigma_V\right)\Omega_\rho^{-\frac{1}{2}}\right)}. \tag{35}$$

To provide intuition about this choice, consider the case where all instruments are very weak, so $\Sigma_V - \Sigma_V\Sigma_X^{-1}\Sigma_V$ is close to zero. In this case it is difficult to detect any departure from the reference model with exogenous X. This leads us to fix a large neighborhood around the reference model where we seek to ensure robustness.

3.3 Implementation

In practice our approach requires several inputs from the researcher. First, one needs to specify a model that is more flexible than the reference model in some dimension. In parametric settings this may consist in including additional covariates, or in allowing for a more

[9] Indeed, $G = (I, 0)$, $\nabla_\beta \log f_{\theta(\eta)}(y, x \mid z) = \frac{1}{\sigma^2}x(y - x'\beta)$, $\nabla_\rho \log f_{\theta(\eta)}(y, x \mid z) = \frac{1}{\sigma^2}(x - \Pi z)(y - x'\beta)$, and $H_{\theta(\eta)} = \frac{1}{\sigma^2}\begin{pmatrix}\Sigma_X & \Sigma_V\\ \Sigma_V & \Sigma_V\end{pmatrix}$, where $\Sigma_X = \Pi\Sigma_Z\Pi' + \Sigma_V$.

[10] Recall that Π is assumed known here. A given choice $\widehat{\Pi}$ will correspond to a particular IV estimator. A more general analysis would include Π in the parameter η of the reference model.

general parametric specification of a density function (e.g., a mixture of two normals instead of a normal distribution). The second input is the distance measure that defines the neighborhood of the reference model, together with the size of that neighborhood. Our choice of ϵ is guided by a model detection error approach. Moreover, as we explained above, in the weighted Euclidean case the choice of weights Ω can be informed by a similar calibration strategy.

To implement the method the researcher needs to compute the score and Hessian of the larger model. In complex models such as structural static or dynamic models this computation will be the main task to implement our approach. Since we focus on smooth models, methods based on numerical derivatives can be used. When the likelihood function is intractable but simulating from the model is feasible, one may use simulation-based approximations to likelihood, score and Hessian (e.g., Fermanian and Salanié, 2004, Kristensen and Shin, 2012). Alternatively, one may construct robust adjustments based on moment functions, as we explain in Appendix C.

4 Semi-parametric models

In this section we consider semi-parametric settings, where the reference model is still parametric but the unknown true model contains a nonparametric component. Our focus is on misspecification of distributional functional forms, and we rely on the Kullback-Leibler divergence to define nonparametric neighborhoods with respect to which we assess robustness.

4.1 Setup and minimum-MSE estimator

Consider a model where the likelihood function has a mixture structure. The distribution of outcomes Y supported on $\mathcal{Y}$ depends on a latent variable A supported on $\mathcal{A}$. We denote the conditional distribution by $g_\beta(y \mid a)$, for β a finite-dimensional parameter. In turn, the distribution of A is denoted by π. The researcher postulates a parametric reference specification for π, which we denote as $\pi_\gamma(a)$, for γ a finite-dimensional parameter. However, she entertains the possibility that her specification may be misspecified in a nonparametric sense. Her goal is to estimate a function of θ_0, $\delta_{\theta_0} = \int \Delta(a, \beta_0)\pi_0(a)da$, which is linear in π_0. In the next subsection we analyze a class of panel data models as one illustration of this setup. In Appendix B we describe two additional examples: a treatment effects model under selection on observables where the conditional mean of potential outcomes may be misspecified, and

a demand model where the distributional assumptions on unobserved preference shocks may be invalid.

In this setup, $\theta = (\beta, \pi)$, $\eta = (\beta, \gamma)$, and $\theta(\eta) = (\beta, \pi_\gamma)$. As a distance measure on θ we use a combination of a weighted Euclidean norm on β and twice the Kullback-Leibler divergence on π; that is, $d(\theta_0, \theta) = \|\beta_0 - \beta\|^2_{\Omega_\beta} + 2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right)\pi_0(a)da$. Neither the choice of Ω_β nor the weighting of the parametric and nonparametric parts play any role in the analysis that follows.[11]

It is instructive to start with the case where both β and γ are assumed to be known. By (17) the small-ϵ approximation to the worst-case bias of an asymptotically linear estimator $\widehat{\delta}$ with influence function h is

$$b_\epsilon(h, \beta, \gamma) = \epsilon^{\frac{1}{2}} \sqrt{\mathrm{Var}_\gamma\left(\Delta(A, \beta) - \mathbb{E}_\beta\left[h(Y) \,|\, A\right]\right)}, \tag{36}$$

where, here and in the following, β, γ, and (β, γ) subscripts indicate that expectations and variances are taken with respect to the joint distribution of the reference model at (β, γ) or some conditional distribution based on it. This bias expression can be used to form confidence intervals for δ_{θ_0}, as explained in Subsection 2.4.

Moreover, by (18), $h_\epsilon^{\mathrm{MMSE}}$ minimizes the following small-ϵ approximation to the MSE,

$$\epsilon\,\mathrm{Var}_\gamma\left(\Delta(A, \beta) - \mathbb{E}_\beta\left[h(Y, \beta, \gamma) \,|\, A\right]\right) + \frac{\mathrm{Var}_{\beta,\gamma}\, h(Y, \beta, \gamma)}{n}, \tag{37}$$

subject to $\mathbb{E}_{\beta,\gamma}\, h(Y, \beta, \gamma) = 0$. The associated first-order conditions are

$$\begin{aligned}\mathbb{E}_{\beta,\gamma}\left[\mathbb{E}_\beta(h_\epsilon^{\mathrm{MMSE}}(Y, \beta, \gamma) \,|\, A) \mid y\right] + (\epsilon n)^{-1} h_\epsilon^{\mathrm{MMSE}}(y, \beta, \gamma) \\ = \mathbb{E}_{\beta,\gamma}\left[\Delta(A, \beta) \mid y\right] - \mathbb{E}_\gamma \Delta(A, \beta), \quad \text{for all } y \in \mathcal{Y},\end{aligned} \tag{38}$$

where the expectations in the terms in brackets are with respect to the posterior distribution $p_{\beta,\gamma}(a \,|\, y) = \frac{g_\beta(y \,|\, a)\pi_\gamma(a)}{\int_{\mathcal{A}} g_\beta(y \,|\, \widetilde{a})\pi_\gamma(\widetilde{a})d\widetilde{a}}$ of the latent variable A given the outcome Y. Note that (38) is linear in $h_\epsilon^{\mathrm{MMSE}}$. This is a Fredholm type-II integral system, which can be solved uniquely given $\epsilon n > 0$, irrespective of the support of Y being finite or infinite. In Subsection 4.3 we describe how we compute the unique minimum-MSE h function in practice.

To provide intuition about the form of $h_\epsilon^{\mathrm{MMSE}}$ it is useful to write the MSE-minimization problem as a functional problem on Hilbert spaces of square-integrable functions. Indeed,

[11] We obtain the same expressions in case the neighborhoods are defined in terms of the KL divergence between *joint* distributions of (Y, A), $\widetilde{d}(\theta_0, \theta) = 2\iint_{\mathcal{Y}\times\mathcal{A}} \log\left(\frac{g_{\beta_0}(y \,|\, a)\pi_0(a)}{g_\beta(y \,|\, a)\pi(a)}\right) g_{\beta_0}(y \,|\, a)\pi_0(a)dyda$, provided $\mathbb{E}_{\beta,\gamma}\left[(\nabla_\beta \log g_\beta(Y \,|\, A))(\nabla_\beta \log g_\beta(Y \,|\, A))'\right]$ is non-singular.

minimizing the MSE is equivalent to minimizing

$$\|\Delta - \delta - \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}\, h\|_{\mathcal{A}}^2 + (\epsilon n)^{-1}\|h\|_{\mathcal{Y}}^2, \tag{39}$$

where $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$ is the conditional expectation operator of Y given A, $\delta = \mathbb{E}_\gamma \Delta(A,\beta)$, $\|g\|_{\mathcal{A}}^2 = \int_{\mathcal{A}} g(a)^2 \pi_\gamma(a) da$, and $\|h\|_{\mathcal{Y}}^2 = \iint_{\mathcal{Y}\times\mathcal{A}} h(y)^2 g_\beta(y\,|\,a)\pi_\gamma(a) dy da$. The unbiasedness constraint on h is automatically satisfied at the solution.

By standard results in functional analysis (e.g., Engl *et al.*, 2000), (39) is minimized at the following *regularized inverse* of the operator $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$ evaluated at $\Delta - \delta$

$$h_\epsilon^{\text{MMSE}} = \left[\mathbb{H}_{\mathcal{Y}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{Y}}\right]^{-1}\left(\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\,\Delta - \delta\right), \tag{40}$$

where $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$ is the conditional expectation operator of A given Y,[12] $\mathbb{I}_{\mathcal{Y}}$ is the identity operator on $\mathcal{Y}$, and $\mathbb{H}_{\mathcal{Y}}$ is the composition of $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$ and $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$; that is,

$$\mathbb{H}_{\mathcal{Y}}[h](y) = \mathbb{E}_{\beta,\gamma}[\,\mathbb{E}_\beta(h(\widetilde{Y})\,|\,A)\,|\,Y = y\,],\ \text{for all } y \in \mathcal{Y}.$$

The function on the right-hand side of (40) is the unique solution to (38). It is well-defined even when $\mathbb{H}_{\mathcal{Y}}$ is singular or its inverse is ill-posed. The term $(\epsilon n)^{-1}$ can be interpreted as a *Tikhonov* penalization (e.g., Carrasco *et al.*, 2007).

Equivalently, (40) can be written as

$$h_\epsilon^{\text{MMSE}} = \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\,\left[\mathbb{H}_{\mathcal{A}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\right]^{-1}(\Delta - \delta), \tag{41}$$

where $\mathbb{I}_{\mathcal{A}}$ is the identity operator on $\mathcal{A}$, and $\mathbb{H}_{\mathcal{A}}$ is the composition of $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$ and $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$. This formula is the semi-parametric counterpart to (29). In Appendix B we describe the mapping between the general setup of Section 2 and the semi-parametric model of this section.

Consider next the case where (β,γ) are estimated. Writing the first-order conditions of (12), and making use of (4), we obtain the following formula for the minimum-MSE h function,

$$\begin{aligned}&\mathbb{Q}_{\beta,\gamma}\mathbb{E}_{\beta,\gamma}\left[\mathbb{E}_\beta(h_\epsilon^{\text{MMSE}}(Y,\beta,\gamma)\,|\,A)\ |\ y\right] + (\epsilon n)^{-1}h_\epsilon^{\text{MMSE}}(y,\beta,\gamma)\\ &\quad = (\epsilon n)^{-1}\left[\nabla_{\beta,\gamma}\log f_{\beta,\pi_\gamma}(y)\right]' H_{\beta,\gamma}^{-1}\,\nabla_{\beta,\gamma}\,\mathbb{E}_\gamma\Delta(A,\beta) + \mathbb{Q}_{\beta,\gamma}\Big(\mathbb{E}_{\beta,\gamma}\left[\Delta(A,\beta)\ |\ y\right] - \mathbb{E}_\gamma\Delta(A,\beta)\Big),\end{aligned} \tag{42}$$

[12] $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$ and $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$ are adjoint operators.

where $\mathbb{Q}_{\beta,\gamma}$ is the operator which projects functions of y onto the orthogonal of the score of the reference model; that is,

$$\mathbb{Q}_{\beta,\gamma}\, h(y) = h(y) - \left[\nabla_{\beta,\gamma} \log f_{\beta,\pi_\gamma}(y)\right]' H_{\beta,\gamma}^{-1} \mathbb{E}_{\beta,\gamma}\left[h(Y)\nabla_{\beta,\gamma} \log f_{\beta,\pi_\gamma}(Y)\right].$$

The system (42) is again linear in h_ϵ^{MMSE}. Note that (42) applies in particular to the case where $\Delta(A,\beta) = \beta_k$ is a component of β.

Finally, to set ϵ we rely on (26). In the case where (β,γ) are known this formula takes the following simple expression

$$\epsilon = \frac{4\left(\Phi^{-1}(p)\right)^2}{n}, \tag{43}$$

which follows from the fact that the maximum eigenvalue of the operator $\mathbb{H}_{\mathcal{A}}$ is equal to one, see Appendix B. Given a detection error probability p we select $\epsilon = \epsilon(p)$ according to (43). When (β,γ) are estimated, the relevant maximal eigenvalue can be approximated numerically, as we describe in Subsection 4.3.

4.2 Application: individual effects in panel data

As a semi-parametric example we study a panel data model with n cross-sectional units and T time periods. For each individual $i = 1,\ldots,n$ we observe a vector of outcomes $Y_i = (Y_{i1},\ldots,Y_{iT})$, and a vector of conditioning variables X_i. The observed data includes both Y's and X's. Observations are i.i.d. across individuals. The distribution of Y_i is modeled conditional on X_i and a vector of latent individual-specific parameters A_i. Leaving i subscripts implicit for conciseness, we denote the corresponding probability density or probability mass function as $g_\beta(y \,|\, a, x)$. In turn, the density of latent individual effects is denoted as $\pi(a \,|\, x)$. The density of Y given X is then

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

The density of X, denoted as f_X, is left unspecified. This setup covers both static models and dynamic panel models, in which case X includes exogenous covariates and initial values of outcomes and predetermined covariates (e.g., Arellano and Bonhomme, 2011).

In panel data settings we are interested in estimating average effects of the form

$$\delta_{\theta_0} = \mathbb{E}_{\theta_0}\left[\Delta(A,X,\beta_0)\right] = \iint_{\mathcal{A}\times\mathcal{X}} \Delta(a,x,\beta_0)\pi_0(a \,|\, x) f_X(x) da dx, \tag{44}$$

for a known function Δ. Average effects, such as average partial effects in static or dynamic discrete choice models, moments of individual effects, or more general policy parameters, are of great interest in panel data applications (Wooldridge, 2010). Since common parameters β_0 can be obtained from (44) by taking $\Delta(A, X, \beta_0) = \beta_{0k}$ for any component of β_0, our framework covers estimation of – and inference on – both average effects and common parameters. The researcher postulates a correlated random-effects specification $\pi_\gamma(a \,|\, x)$ indexed by a parameter γ. For example, a common specification in applied work is a normal distribution whose mean depends linearly on X's and whose variance is constant (Chamberlain, 1984).

Random-effects and empirical Bayes estimators. In the next section we will compare the finite-sample performance of the minimum-MSE estimator of δ_{θ_0}, obtained by minimizing the worst-case conditional MSE given covariates $X_1, \ldots, X_n$, to that of two other commonly used panel data estimators. The first one is the *random-effects* (RE) estimator

$$\widehat{\delta}^{\mathrm{RE}} = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{A}} \Delta\left(a, X_i, \widehat{\beta}\right) \pi_{\widehat{\gamma}}(a \,|\, X_i)\, da, \tag{45}$$

where $(\widehat{\beta}, \widehat{\gamma})$ is the MLE of (β, γ) based on the reference model. The second one is the *empirical Bayes* (EB) estimator

$$\widehat{\delta}^{\mathrm{EB}} = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{A}} \Delta(a, X_i, \widehat{\beta}) \underbrace{\frac{g_{\widehat{\beta}}(Y_i \,|\, a, X_i)\pi_{\widehat{\gamma}}(a \,|\, X_i)}{\int_{\mathcal{A}} g_{\widehat{\beta}}(Y_i \,|\, \widetilde{a}, X_i)\pi_{\widehat{\gamma}}(\widetilde{a} \,|\, X_i) d\widetilde{a}}}_{= p_{\widehat{\beta},\widehat{\gamma}}(a \,|\, Y_i, X_i)} da, \tag{46}$$

where $p_{\widehat{\beta},\widehat{\gamma}}(a \,|\, Y_i, X_i)$ is the posterior distribution of A_i given (Y_i, X_i) implied by $g_{\widehat{\beta}}$ and $\pi_{\widehat{\gamma}}$. Both $\widehat{\delta}^{\mathrm{RE}}$ and $\widehat{\delta}^{\mathrm{EB}}$ are consistent for fixed T as n tends to infinity under correct specification of the reference model. Our interest centers on situations where misspecification of π_γ makes such commonly used estimators fixed-T inconsistent. Settings where g_β is assumed correctly specified while π_γ may be misspecified have received substantial attention in the panel data literature (e.g., Heckman, 1981, Arellano and Hahn, 2007).[13]

Our approach allows us to rank the RE and EB estimators in terms of bias. For simplicity here we focus on β and γ being known. The small-ϵ approximation to the bias of the RE estimator is

$$b_\epsilon(h^{\mathrm{RE}}, \beta, \gamma) = \epsilon^{\frac{1}{2}} \sqrt{\widehat{\mathrm{Var}}_\gamma\left(\Delta(A, X, \beta)\right)},$$

[13] Our approach allows us to consider other forms of model misspecification than the sole misspecification of the distribution of individual effects. In Appendix B we provide additional results where either g_β, or both g_β and π_γ, are misspecified.

where $\widehat{\mathrm{Var}}_\gamma\left(\Delta(A,X,\beta)\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\gamma\,|\,X_i}\left(\Delta(A,X_i,\beta)\right)$, with the variance being computed with respect to the conditional density $\pi_\gamma(\cdot\,|\,X_i)$. The corresponding bias expression for the EB estimator is

$$b_\epsilon(h^{\mathrm{EB}},\beta,\gamma) = \epsilon^{\frac{1}{2}}\sqrt{\widehat{\mathrm{Var}}_\gamma\left(\Delta(A,X,\beta) - \mathbb{E}_\beta\left[\mathbb{E}_{\beta,\gamma}\left(\Delta(\widetilde{A},X,\beta)\,|\,Y,X\right)\,|\,A,X\right]\right)},$$

where $\widetilde{A}$ has the same distribution as A given Y,X. It thus follows that $b_\epsilon(h^{\mathrm{EB}},\beta,\gamma) \leq b_\epsilon(h^{\mathrm{RE}},\beta,\gamma)$. Hence, from a fixed-$T$ robustness perspective, the EB estimator dominates the RE estimator in terms of bias. In addition, as T tends to infinity we expect $b_\epsilon(h^{\mathrm{EB}},\beta,\gamma)$ to tend to zero.[14] By contrast, $b_\epsilon(h^{\mathrm{RE}},\beta,\gamma)$ is constant, independent of T. This comparison is in line with the consistency of EB estimators and *in*consistency of RE estimators of average effects under large T (Arellano and Bonhomme, 2009).

Link to the semi-parametric panel data literature. Similarly to the EB estimator, but unlike the RE estimator, the minimum-MSE estimator updates the prior π_γ in light of the data. However, the form of the minimum-MSE update rule in (42) differs from Bayesian updating. Here we relate our estimator to the semi-parametric panel data literature. In Appendix C we discuss the link between our approach and Bayesian approaches.

Consider first the estimation of the average of $\Delta(A,\beta)$, assuming (β,γ) known. (40) shows that in this case the minimum-MSE h function is obtained by Tikhonov regularization. It is well-understood that average effects are typically only partially identified in discrete choice panel data models (Chernozhukov *et al.*, 2013, Pakes and Porter, 2013), and that in point-identified models with continuous outcomes they may not be root-n estimable due to ill-posedness (Bonhomme and Davezies, 2017). The presence of the term $(\epsilon n)^{-1}$ in (40), which is constant in large samples under our calibration, bypasses these issues by making the operator $[\mathbb{H}_{\mathcal{Y}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{Y}}]$ non-singular.

In some cases, and under strong conditions, regular estimation of average effects is possible provided there exists a suitable function ζ such that $\mathbb{E}_\beta\left[\zeta(Y,X)\,|\,A=a,X=x\right] = \Delta(a,x,\beta)$. In such cases it follows from (40) that $\lim_{\epsilon\to\infty}\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \frac{1}{n}\sum_{i=1}^n \zeta(Y_i,X_i)$, so in the large-$\epsilon$ limit the minimum-MSE estimator remains unbiased for δ_{θ_0} under misspecifi-

[14] This is easy to see in a model without covariates X since, as T tends to infinity, we expect that

$$\mathbb{E}_\beta\left[\mathbb{E}_{\beta,\gamma}\left(\Delta(\widetilde{A},\beta)\,|\,Y\right)\,|\,A=a\right] \approx \mathbb{E}\left(\Delta(\widehat{A}(Y,\beta),\beta)\,|\,A=a\right) \approx \Delta(a,\beta),\ \text{for all } a,$$

where $\widehat{A}(y,\beta) = \mathrm{argmax}_a\ g_\beta(y\,|\,a)$ is the maximum likelihood estimator of A (for a given individual).

cation (for known β, γ). More generally, by focusing on a shrinking neighborhood of the distribution π_γ, as opposed to entertaining any possible distribution, our approach avoids issues of non-identification and ill-posedness while guaranteeing MSE-optimality within that neighborhood.

Next, consider the estimation of $c'\beta$, for c a $\dim \beta \times 1$ vector and γ known. Note that, by the Woodbury identity,

$$(\epsilon n)^{-1}\left[\mathbb{H}_{\mathcal{Y}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{Y}}\right]^{-1} = \mathbb{I}_{\mathcal{Y}} - \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\left[\mathbb{H}_{\mathcal{A}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\right]^{-1}\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} := \mathbb{W}^{\epsilon}_{\beta,\pi_\gamma}$$

is a regularized counterpart to the functional differencing "within" projection operator (Bonhomme, 2012). It then follows from (42) that

$$h^{\text{MMSE}}_{\epsilon}(y,\beta,\gamma) = \mathbb{W}^{\epsilon}_{\beta,\pi_\gamma}[\nabla_\beta \log f_{\beta,\pi_\gamma}](y)'\left\{\mathbb{E}_{\beta,\gamma}\left(\nabla_\beta \log f_{\beta,\pi_\gamma}(Y)\mathbb{W}^{\epsilon}_{\beta,\pi_\gamma}[\nabla_\beta \log f_{\beta,\pi_\gamma}](Y)'\right)\right\}^{-1} c. \tag{47}$$

As ϵ tends to infinity, $\mathbb{W}^{\epsilon}_{\beta,\pi_\gamma}$ tends to the functional differencing projection operator $\mathbb{W}_{\beta,\pi_\gamma} = \mathbb{I}_{\mathcal{Y}} - \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\mathbb{E}^{\dagger}_{\mathcal{A}\,|\,\mathcal{Y}}$, where $\mathbb{E}^{\dagger}_{\mathcal{A}\,|\,\mathcal{Y}}$ denotes the Moore-Penrose generalized inverse of $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$. In this limit, the minimum-MSE estimator is the one-step approximation to the semi-parametric efficient estimator of $c'\beta_0$ based on the efficient score $\mathbb{W}_{\beta_0,\pi_0}[\nabla_\beta \log f_{\beta_0,\pi_0}](y)$.

Yet, the efficient estimator fails to exist when the matrix denominator in (47) is singular. For example, in discrete choice models common parameters are generally not point-identified (Chamberlain, 2010, Honoré and Tamer, 2006). In models with continuous outcomes, identification and regularity require high-level *non-surjectivity* conditions (related to so-called "completeness" conditions) which may be hard to verify. Here the term $(\epsilon n)^{-1}$ acts as a regularization of the functional differencing projection. Compared to semi-parametric estimation based on functional differencing, the approach described in this section covers a large class of models with continuous or discrete outcomes, and it does not require optimization. In the next section we will see that our estimator provides reliable results in simulations of a dynamic probit model, in settings with a substantial amount of misspecification.

4.3 Implementation

Unlike for the parametric models we studied in Section 3, the minimum-MSE h function is generally not available in closed form in semi-parametric models. Here we describe how we compute a numerical approximation to the minimum-MSE estimator $\widehat{\delta}^{\text{MMSE}}_{\epsilon} = \mathbb{E}_{\widetilde{\gamma}}\Delta(A,\widetilde{\beta}) + \frac{1}{n}\sum_{i=1}^{n} h^{\text{MMSE}}_{\epsilon}(Y_i,\widetilde{\beta},\widetilde{\gamma})$, where $h^{\text{MMSE}}_{\epsilon}$ is given by (42) and $(\widetilde{\beta},\widetilde{\gamma})$ are preliminary estimates.

We abstract from conditioning covariates. In the presence of covariates X_i we use the same technique to approximate $h_\epsilon^{\text{MMSE}}(\cdot \,|\, x)$ for each value of $X_i = x$. We use this approach in the numerical illustration on a dynamic panel data model in the next section, where the covariate is the initial condition.

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 $\Delta(A^{(s)}, \beta)$, 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 (42), a fixed-S approximation to the minimum-MSE estimator is then

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

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

In turn, $\epsilon(p)$ can be approximated as $4\Phi^{-1}(p)^2/(n\lambda_{max})$, where λ_{max} is the maximum eigenvalue of $G'QG = \widetilde{G}'\widetilde{G}$, see Appendix B. In the known (β, γ) case, $\lambda_{max} = 1$ since it is the maximal eigenvalue of the stochastic matrix $G'G$. In practice, when (β, γ) are estimated and low-dimensional, λ_{max} appears to be often close to one.

5 Revisiting the dynamic panel probit model

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

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

where $U_{i1}, ..., U_{iT}$ are i.i.d. standard normal, independent of A_i and Y_{i0}. Here Y_{i0} is observed, so there are effectively $T+1$ time periods. We focus on the average *state dependence* effect

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

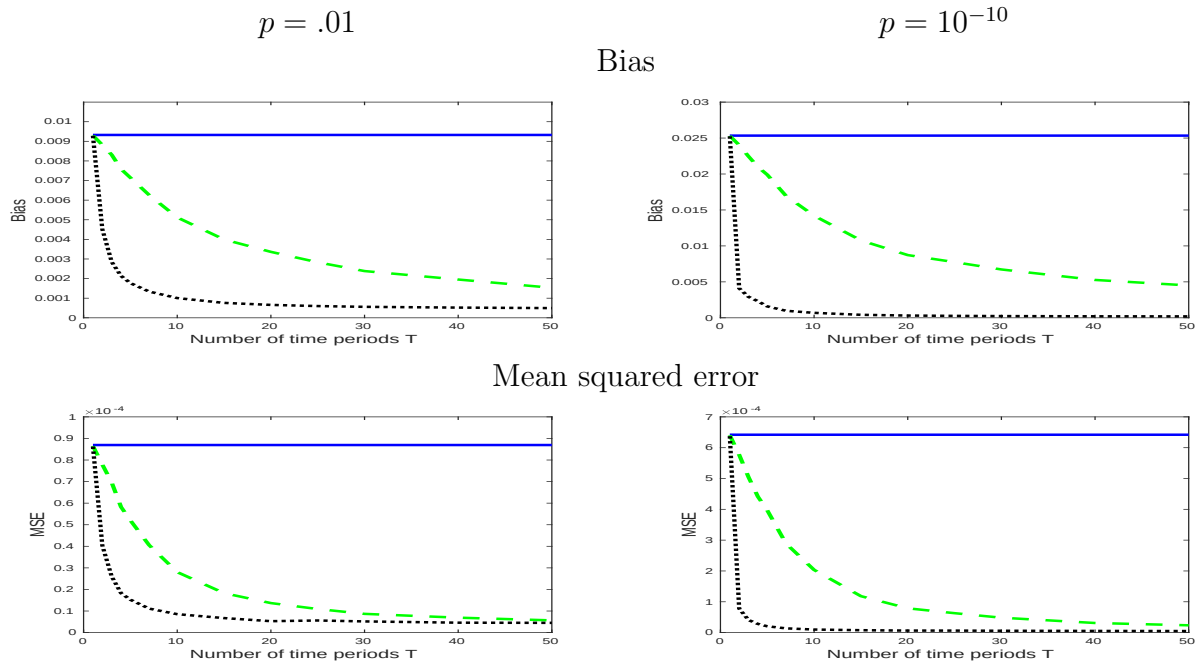
We assume that the probit conditional likelihood given individual effects and lagged outcomes is correctly specified. However we do not assume knowledge of π_0 or its functional form. We specify the reference density $\pi = \pi_{\mu,\sigma}$ of A_i given Y_{i0} as a Gaussian with mean $\mu_1 + \mu_2 Y_{i0}$ and variance σ^2. Throughout this section we treat the parameters (μ, σ) of the reference model as known.

The dynamic probit model has proven challenging to analyze. No semi-parametrically consistent estimators of β and δ are available in the literature. Moreover, it has been documented that static and dynamic probit models are typically partially identified for fixed T (Chamberlain, 2010, Honoré and Tamer, 2006). Here we report simulation results suggesting that our minimum-MSE estimator can perform well under sizable departures from the reference model.

Before showing results on simulated data we start by reporting illustrative calculations of the bias of the random-effects (RE), empirical Bayes (EB), and minimum-MSE estimators. To do so we set $\beta_0 = .5$, $\mu_1 = -.25$, $\mu_2 = .5$, and $\sigma = .8$. In the calculations we treat (β_0, μ, σ) as known, and evaluate the formulas at $Y_{i0} = 0$. To compute the bias of the minimum-MSE estimator we use the approach described in Subsection 4.3, based on $S = 30,000$ simulated draws. We use a similar approach to compute the bias of RE and EB estimators. We vary T between 1 and 50. The variance and MSE formulas are calculated at a sample size $n = 500$. In Figure 1 we show the asymptotic bias b_ϵ and MSE for each of the three estimators, where ϵ is set according to (43) for a detection error probability $p = .01$ (left graph) and $p = 10^{-10}$ (right), and a sample size $n = 500$.

On the top panel of Figure 1 we see that the bias of the RE estimator (solid line) is the largest, and that it does not decrease as T grows. By contrast, the bias of the EB estimator (dashed) decreases as T grows. Interestingly, the bias of the minimum-MSE estimator (dotted) is the smallest, and it decreases quickly as T increases. The bias levels off in the large-T limit, since ϵ is indexed by n and independent of T. Setting p to the much smaller value $p = 10^{-10}$ implies larger biases for the RE and EB estimators. Lastly, on the bottom panel we observe a similar relative ranking between estimators in terms of MSE.

Figure 1: Bias and MSE of different estimators of the average state dependence effect in the dynamic probit model

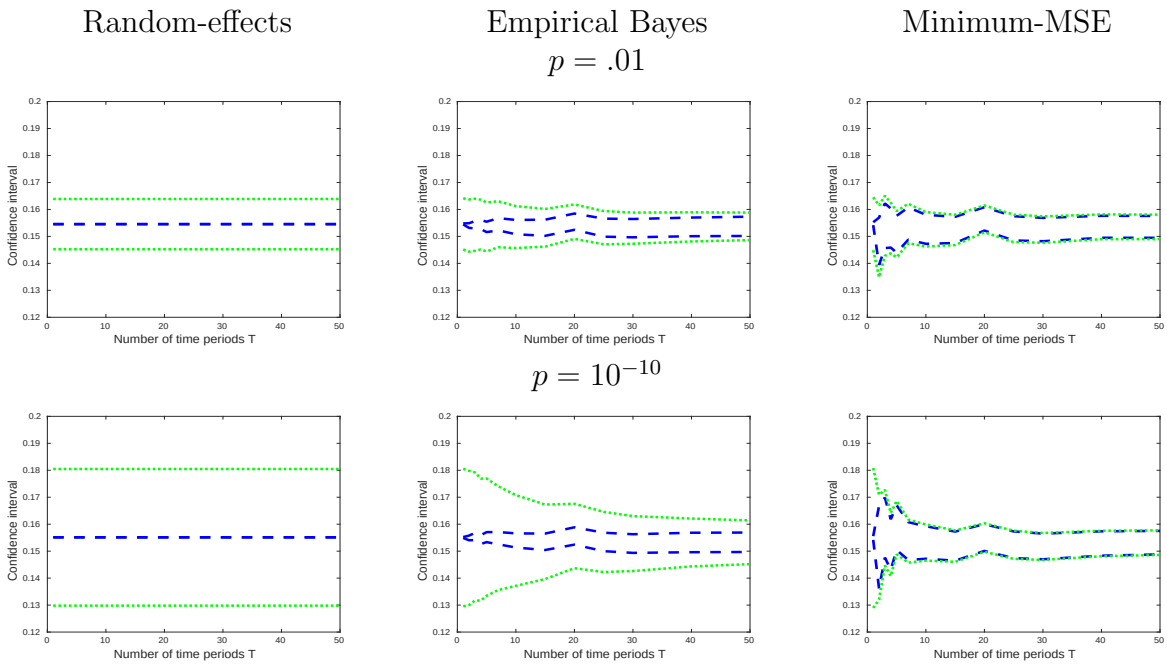


Notes: Asymptotic bias b_ϵ (top panel) and MSE (bottom panel) for different panel length T. The solid line corresponds to the random-effects estimator $\widehat{\delta}^{\text{RE}}$, the dashed line to the empirical Bayes estimator $\widehat{\delta}^{\text{EB}}$, and the dotted line to the minimum-MSE estimator $\widehat{\delta}^{\text{MMSE}}_{\epsilon}$. ϵ is chosen according to (43) for a detection error probability $p = .01$ (left) and $p = 10^{-10}$ (right) when $n = 500$. (β_0, μ, σ) are treated as known.

We then turn to calculations of confidence intervals. In Figure 2 we report two types of asymptotic 95% confidence intervals for the average state dependence effect: obtained under correct specification (dashed lines), and allowing for local misspecification as in (22) (dotted lines).[15] ϵ is chosen based on (43) for a probability $p = .01$ (top panel) and $p = 10^{-10}$ (bottom), and a sample size $n = 500$. We see that accounting for model misspecification leads to enlarged confidence intervals. However the size of the enlargement varies to a large extent with the estimator considered, reflecting the amount of bias. In particular, the confidence intervals based on the minimum-MSE estimator are quite similar under correct specification and misspecification. Moreover, while for $p = 10^{-10}$ the confidence intervals based on the RE and EB estimators widen substantially, those based on the minimum-MSE estimator remain quite informative.

[15] We also computed the Armstrong and Kolesár (2016) confidence intervals in this case, see footnote 6. Those are almost identical to the ones we report in Figure 2.

Figure 2: Confidence intervals of the average state dependence effect in the dynamic probit model



Notes: Asymptotic 95%-confidence intervals for the average state dependence effect, based on three estimators. Dashed lines correspond to confidence intervals based on correct specification, dotted lines to the ones allowing for local misspecification. $n = 500$. ϵ *is chosen according to (43) for a detection error probability* $p = .01$ *(top) and* $p = 10^{-10}$ *(bottom).* (β_0, μ, σ) *are treated as known.*

Monte Carlo simulations. We next report the results of two Monte Carlo simulations under misspecification. In the first one we use the same data generating process as above, except that we set the population distribution of A_i to be log-normal with mean $-.25+.5Y_{i0}$ and standard deviation .8. The assumed distribution for A_i in the parametric reference model is still Gaussian, with the same mean and standard deviation. Here we estimate β_0 along with the average state dependence effect δ_{θ_0}, and treat the parameters (μ, σ) of the reference model as fixed. $\beta_0 = .5$, and Y_{i0} are drawn from a Bernoulli(1/2). We use $S = 1000$ simulated draws to compute the estimators. Our goal is to document the performance of the minimum-MSE estimator under a particular form of global misspecification.

In Table 1 we report the results of 1000 Monte Carlo replications, for T ranging between 5 and 50, and $n = 500$. The upper panel shows the bias and MSE for the average state dependence effect δ, and the lower panel shows the bias and MSE for the autoregressive parameter β. We report results for δ for five estimators: the RE estimator, the EB estimator, the linear probability (LP) estimator, and the minimum-MSE estimators with ϵ set according

Table 1: Monte Carlo simulation of the average state dependence effect and autoregressive parameter in the dynamic probit model, DGP with log-normal A_i

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

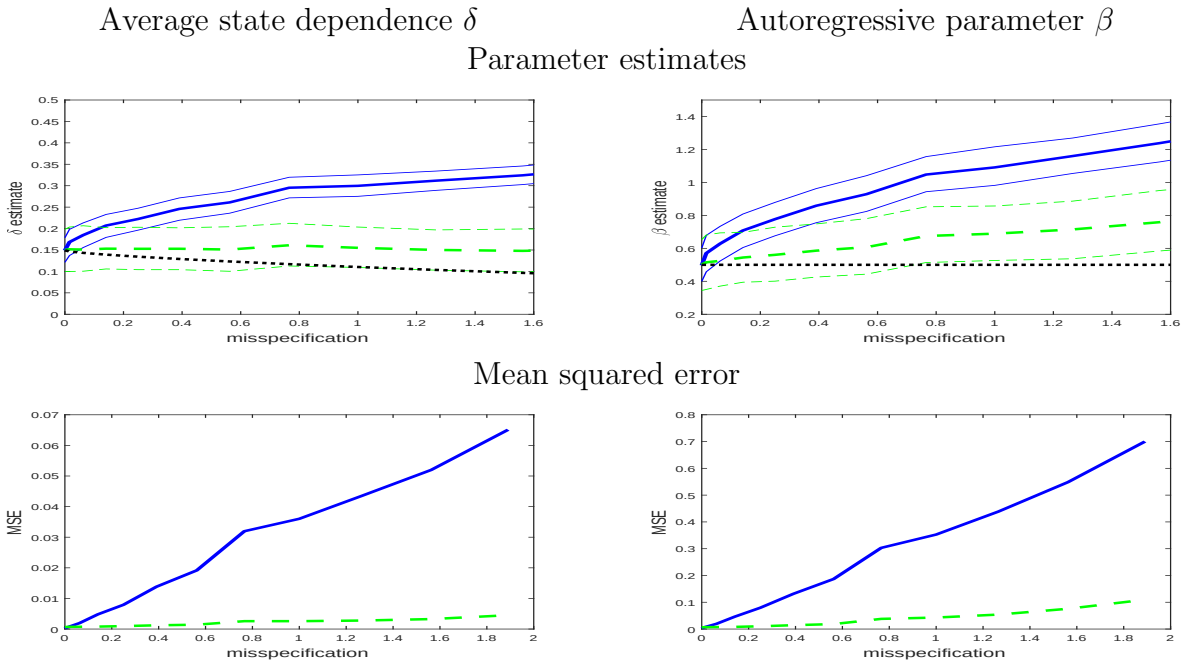
Notes: $n = 500$*, results for* 1000 *simulations.* (μ, σ) *are treated as known.*

to $p = .01$ and $p = 10^{-10}$, respectively. We report results for β for three estimators: the random-effects MLE and the two minimum-MSE estimators.

Focusing first on β, we see that the MLE is biased due to the misspecification of the random-effects density. When $T = 5$ the mean estimate is .35, compared to a true value of .5. Both minimum-MSE estimators reduce the bias substantially, the mean estimates being .46 and .49 depending on the value of ϵ. This bias reduction comes with some increase in variance: for example when $T = 5$ the variance of the minimum-MSE estimator for ϵ calibrated to $p = .01$ is .0066 compared to .0026 for the MLE. Yet the overall mean squared error is lower for our robust estimator compared to the MLE. Turning to average state dependence δ, we see that the RE estimator is substantially biased and has large MSE in this case too. In comparison the EB estimator has smaller bias and mean squared error. The LP estimator is severely biased in short panels in this dynamic setting. The minimum-MSE estimator performs again clearly best in terms of both bias and MSE.

Note that in this simulation design the calibrated neighborhood size ϵ is .04 for $p = .01$, and .32 for $p = 10^{-10}$, whereas twice the Kullback-Leibler divergence between the true log-normal density and the assumed normal density is equal to 1.52. Hence the true distribution of A_i lies quite far outside of the chosen neighborhood. It represents a form of misspecification that should be "easy to detect" from the reference model at conventional significance levels.

Figure 3: Estimates and mean squared error of random-effects and minimum-MSE estimators under varying amount of misspecification



Notes: Random-effects (solid) and minimum-MSE (dashed) for δ (left graphs) and β (right graphs). True parameter values are shown in dotted. $n = 500$, $T = 5$. The reference specification for π is normal with mean $-.25 + .5Y_{i0}$ and standard deviation .8, whereas the true π_0 is normal with the same standard deviation and mean $-.25+\nu+.5Y_{i0}$. On the x-axis we report twice the KL divergence; that is, $\nu^2/.64$. Top panel: mean and 95% interval. Bottom panel: mean squared error. ϵ is chosen according to (43) for a detection error probability $p = .01$. (μ, σ) are treated as known.

In spite of this, the minimum-MSE estimator provides effective bias and MSE reduction in this environment.

In order to better understand the sensitivity of our estimator to misspecification, we perform a second simulation where we vary the amount of misspecification. While the reference specification for π is still normal with mean $-.25 + .5Y_{i0}$ and standard deviation .8, the true π_0 is now normal with mean $-.25 + \nu + .5Y_{i0}$ and the same standard deviation. On the x-axes in Figure 3 we report twice the Kullback-Leibler divergence between π_0 and π; that is, $\nu^2/.64$. Hence the amount of misspecification increases as one moves to the right. The RE estimator is shown in solid in the graphs, whereas the minimum-MSE estimator for $p = .01$ is shown in dashed, and the true parameter value is in dotted. We see that, for both δ and β, the minimum-MSE estimator is less sensitive to departure from correct

specification than the RE estimator. Although this robustness comes at a price in terms of variance under correct specification (that is, when $\nu = 0$) the comparison of bias and MSE clearly favors our estimator as soon as some misspecification is allowed for. The results for ϵ calibrated to $p = 10^{-10}$ can be found in the appendix.

6 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.

6.1 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

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

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.[16] 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

[16] 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.

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 $\theta = (a, b, d, \lambda)$, $\eta = (a, b, d)$, $\theta(\eta) = (a, b, d, 0)$ and $\delta_\theta = \mathbb{E}\left(\Phi\left[a(Y+W) - (a+d)(Y+\tau^{\text{treat}}) - \lambda\tau^{\text{treat}} - b\right] - \Phi\left[a(Y+W) - (a+d)Y - b\right]\right)$. We focus on the effect on eligible (i.e., poorer) households. We will first estimate $\delta_{\theta(\eta)}$ using the control villages only. We will then compute the minimum-MSE estimator $\widehat{\delta}_\epsilon^{\text{MMSE}}$, for given $\epsilon = \epsilon(p)$, 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.

6.2 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.

In Table 2 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 calibrate ϵ based on a detection error probability $p = .01$. 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 2 we report estimates of the minimum-MSE estimator for the same ϵ,

Table 2: 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	.053	.087	.050
non-robust CI	(.006,.147)	(.032,.129)	-	-	-	-
robust CI	(-.053,.205)	(-.062,.222)	(-.012,.166)	(-.023,.129)	-	-
	Counterfactual 1: doubling subsidy					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.145	.146	.146	.104	-	-
robust CI	(-.085,.374)	(-.085,.378)	(-.012,.304)	(-.019,.227)	-	-
	Counterfactual 2: unconditional transfer					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.004	.005	.004	-.018	-	-
robust CI	(-.585,.593)	(-.486,.497)	(-.252,.260)	(-.238,.201)	-	-

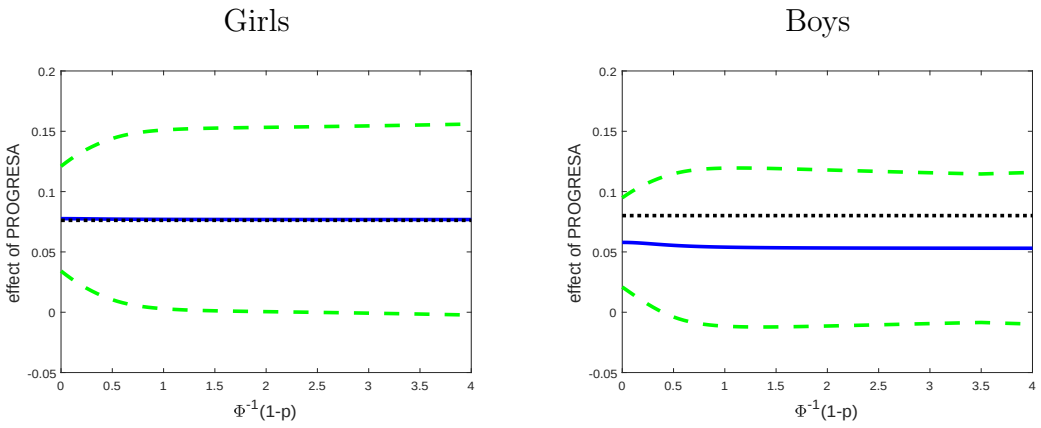
Notes: Sample from Todd and Wolpin (2006). $p = .01$. CI are 95% confidence intervals. The unconditional transfer amounts to 5000 pesos in a year.

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 confidence intervals are then substantially reduced, although they are still large.[17] 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.

[17]The PROGRESA impacts are significant at the 10% level for girls, though not for boys.

Figure 4: Effect of the PROGRESA subsidy as a function of the detection error probability



Notes: Sample from Todd and Wolpin (2006). $\epsilon(p)$ is chosen according to (32), with $\Phi^{-1}(1-p)$ 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. Girls (left) and boys (right).

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 4 we plot the estimates for girls (left) and boys (right) as a function of $\Phi^{-1}(1-p)$, in addition to 95% confidence intervals based on those estimates, where $\epsilon = \epsilon(p)$ is chosen according to (32). In dotted we show the unadjusted model-based predictions. The estimates and confidence intervals reported in Table 2 correspond to $\Phi^{-1}(.99) = 2.32$. 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 .053 for our calibrated 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 2 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.[18] By contrast, we find no effect of an unconditional income transfer.

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. We report the results of this exercise in the appendix.

7 Extensions

In Appendix C we describe several extensions to our approach. In particular, we consider settings where a finite-dimensional parameter θ_0 does not fully determine the distribution f_0 of Y, but satisfies a finite-dimensional system of moment conditions

$$\mathbb{E}_{f_0}\Psi(Y,\theta_0)=0. \tag{48}$$

We focus on asymptotically linear generalized method-of-moments (GMM) estimators of δ_{θ_0} that satisfy

$$\widehat{\delta}=\delta_{\theta(\eta)}+a(\eta)'\frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i,\theta(\eta))+o_P(\epsilon^{\frac{1}{2}})+o_P(n^{-\frac{1}{2}}), \tag{49}$$

for an η-specific parameter vector $a(\eta)$. We characterize the form of $a(\eta)$ which leads to minimum worst-case MSE in Γ_ϵ. We use this framework to revisit the OLS/IV example of Subsection 3.2, removing the Gaussian assumptions on the distributions.

Lastly, in Appendix C we present other extensions regarding different distance or loss functions, a different rule for the neighborhood size ϵ, and the role of the unbiasedness constraint (2). In addition, we discuss how our approach and Bayesian approaches relate to each other, give a result on fixed-ϵ bias in a particular case, and provide a characterization which links our local approach to partial identification.

8 Conclusion

We propose a framework for estimation and inference in the presence of model misspecification. The methods we develop allow one to perform sensitivity analysis for existing estimators, and to construct improved estimators and confidence intervals that are less sensitive to model assumptions.

[18] The estimates are significant at the 10% level for both genders.

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] Anderson, S. P., A. De Palma, and J. F. Thisse (1992): *Discrete Choice Theory of Product Differentiation*. MIT press.

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

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

[4] 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.

[5] Arellano, M., and S. Bonhomme, S. (2009): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77(2), 489–536.

[6] Arellano, M., and S. Bonhomme (2011): "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3(1), 395–424.

[7] Arellano, M., and J. Hahn (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments,". In: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics, Ninth World Congress*, Cambridge University Press.

[8] 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.

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

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

[11] 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.

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

[13] 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.

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

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

[16] 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.

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

[18] Carrasco, M., J. P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," *Handbook of Econometrics*, 6, 5633–5751.

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

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

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

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

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

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

[25] 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.

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

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

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

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

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

[31] Fermanian, J. D., and B. Salanié (2004): "A Nonparametric Simulated Maximum Likelihood Estimation Method," *Econometric Theory*, 20(4), 701–734.

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

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

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

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

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

[37] Hahn, J. and W.K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", *Econometrica*, 72, 1295–1319.

[38] 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.

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

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

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

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

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

[44] 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.

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

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

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

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

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

[50] Kristensen, D., and Y. Shin (2012): "Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood," *Journal of Econometrics*, 167(1), 76–94.

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

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

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

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

[55] Mukhin, Y. (2018): "Sensitivity of Regular Estimators." arXiv preprint arXiv:1805.08883.

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

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

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

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

[60] 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.

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

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

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

[64] Rosenbaum, P. R., and D. B. Rubin (1983a): "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.

[65] Rosenbaum, P. R., and D. B. Rubin (1983b): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

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

[67] Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 267–288.

[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. (2000): *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.

[75] Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

APPENDIX

A Main results

In this section of the appendix we provide the proofs for the main results of Section 2. At the end of the section we give some background on asymptotically linear estimators.

A.1 Proof of Theorem 1

A.1.1 Notation and assumptions

In all our applications Θ is either a vector space or an affine space. Let $T(\Theta)$ and $T^*(\Theta)$ be the tangent and co-tangent spaces of Θ.[19] Thus, for $\theta_1, \theta_2 \in \Theta$ we have $(\theta_1 - \theta_2) \in T(\Theta)$, and $T^*(\Theta)$ is the set of linear maps $u : T(\Theta) \to \mathbb{R}$. For $v \in T(\Theta)$ and $u \in T^*(\Theta)$ we use the bracket notation $\langle v, u \rangle \in \mathbb{R}$ to denote their scalar product. Our squared distance measure $d(\theta_0, \theta(\eta))$ on Θ induces a norm on the tangent space $T(\Theta)$, namely for $v \in T(\Theta)$,

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

For every $\eta \in \mathcal{B}$ we assume that there exists a map $\Omega_\eta : T(\Theta) \to T^*(\Theta)$ such that, for all $v \in T(\Theta)$,

$$\|v\|^2_{\mathrm{ind},\eta} = \langle v, \Omega_\eta v \rangle .$$

We assume that Ω_η is invertible, and write $\Omega_\eta^{-1} : T^*(\Theta) \to T(\Theta)$ for its inverse.

For a scalar function on Θ, such as $\delta : \Theta \mapsto \mathbb{R}$, we have $\nabla_\theta \delta_\theta \in T^*(\Theta)$; that is, the typical element of $T^*(\Theta)$ is a gradient. Conversely, for a map to Θ, such as $\eta \mapsto \theta(\eta)$, we have $\frac{\partial \theta(\eta)}{\partial \eta_k} \in T(\Theta)$. The two versions of the Jacobian $G'_\eta : \mathbb{R}^{\dim \eta} \to T(\Theta)$ and $G_\eta : T^*(\Theta) \to \mathbb{R}^{\dim \eta}$ are defined by

$$G'_\eta : q \mapsto \sum_{k=1}^{\dim \eta} q_k \frac{\partial \theta(\eta)}{\partial \eta_k}, \qquad G_\eta : u \mapsto \left(\left\langle \frac{\partial \theta(\eta)}{\partial \eta_k}, u \right\rangle \right)_{k=1,\ldots,\dim \eta},$$

where $q \in \mathbb{R}^{\dim \eta}$ and $u \in T^*(\Theta)$. Similarly, the Hessian $H_{\theta(\eta)} : T(\Theta) \to T^*(\Theta)$ is defined by

$$H_{\theta(\eta)} : v \mapsto \mathbb{E}_{\theta(\eta)} \left[\left\langle v, \nabla_\theta \log f_{\theta(\eta)}(Y) \right\rangle \nabla_\theta \log f_{\theta(\eta)}(Y) \right].$$

[19] If Θ is a more general manifold (not just an affine space), then the tangent and co-tangent spaces depend on the particular value of $\theta \in \Theta$. We then need a connection on the manifold that provides a map between the tangent spaces at $\theta(\eta)$ and $\theta_0 \in \Gamma_\epsilon(\eta)$. 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.

The definitions of the projected Hessian $\widetilde{H}_{\theta(\eta)} : T(\Theta) \to T^*(\Theta)$ and the projected gradient operator $\widetilde{\nabla}_\theta$ are then as in the main text, namely $\widetilde{H}_{\theta(\eta)} = H_{\theta(\eta)} - H_{\theta(\eta)} G'_\eta H_\eta^{-1} G_\eta H_{\theta(\eta)}$, and $\widetilde{\nabla}_\theta = \nabla_\theta - H_{\theta(\eta)} G'_\eta H_\eta^{-1} \nabla_\eta$. We have $\widetilde{\nabla}_\theta \delta_{\theta(\eta)} \in T^*(\Theta)$.

The dual norm for $u \in T^*(\Theta)$ was defined in the main text. We have

$$\|u\|_\eta = \sup_{v \in T(\Theta)\setminus\{0\}} \frac{\langle v, u \rangle}{\|v\|_{\mathrm{ind},\eta}}, \qquad\qquad \|u\|_\eta^2 = \left\langle \Omega_\eta^{-1} u, u \right\rangle .$$

$\|\cdot\|_\eta$ is also the norm on $T^*(\Theta)$ that is naturally induced by $d(\theta_0, \theta(\eta))$. We use the shorter notation $\|\cdot\|_\eta$ for that norm, because it also appears in the main text. Notice also that $\|\cdot\|_{\mathrm{ind},\eta}$, $\|\cdot\|_\eta$, Ω_η, and Ω_η^{-1} could all be defined for general $\theta \in \Theta$, but since we use them only at the reference values $\theta = \theta(\eta)$ we index them simply by η.

Throughout we assume that $\dim \eta$ is finite. For vectors $w \in \mathbb{R}^{\dim \eta}$ we use the standard Euclidean norm $\|w\|$, and for $\dim \eta \times \dim \eta$ matrices we use the spectral matrix norm, which we also denote by $\|\cdot\|$.

The vector norms $\|\cdot\|_{\mathrm{ind},\eta}$, $\|\cdot\|_\eta$, $\|.\|$ on $T(\Theta)$, $T^*(\Theta)$, $\mathbb{R}^{\dim \eta}$ immediately induce norms on any maps between $T(\Theta)$, $T^*(\Theta)$, $\mathbb{R}^{\dim \eta}$, and $\mathbb{R}$. With a slight abuse of notation we denote all those norms simply by $\|.\|_\eta$. For example, for $H_{\theta(\eta)} : T(\Theta) \to T^*(\Theta)$ we have

$$\left\| H_{\theta(\eta)} \right\|_\eta := \sup_{v \in T(\Theta)\setminus\{0\}} \frac{\|H_{\theta(\eta)} v\|_\eta}{\|v\|_{\mathrm{ind},\eta}} = \sup_{v,w \in T(\Theta)\setminus\{0\}} \frac{\left\langle w, H_{\theta(\eta)} v \right\rangle}{\|v\|_{\mathrm{ind},\eta} \|w\|_{\mathrm{ind},\eta}}.$$

Our first set of assumptions is as follows.

Assumption A1. *We assume that $Y_i \sim$ i.i.d.f_{θ_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) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| \delta_{\theta_0} - \delta_{\theta(\eta)} - \left\langle \theta_0 - \theta(\eta), \nabla_\theta \delta_{\theta(\eta)} \right\rangle \right| = o(\epsilon^{1/2})$.

(iii)
$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right]^2 dy = o(1),$$
$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left\| \nabla_\theta \log f_{\theta(\eta)}(y) \right\|_\eta^2 \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right]^2 dy = o(1),$$
$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y) \right\rangle \right]^2 dy = o(\epsilon).$$

(iv) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \epsilon^{-1/2} \|\theta_0 - \theta(\eta)\|_{\mathrm{ind},\eta} = 1 + o(1)$. *Furthermore, for $u(\eta) \in T^*(\Theta)$ with $\sup_{\eta\in\mathcal{B}} \|u(\eta)\|_\eta = O(1)$ we have*

$$\sup_{\eta\in\mathcal{B}} \left| \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \epsilon^{-1/2} \langle \theta_0 - \theta(\eta), u(\eta) \rangle - \|u(\eta)\|_\eta \right| = o(1).$$

(v) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \|\nabla_\theta \delta_{\theta_0}\|_\eta = O(1)$, $\sup_{\eta\in\mathcal{B}} \left\|H_\eta^{-1}\right\| = O(1)$, $\sup_{\eta\in\mathcal{B}} \|G_\eta\|_\eta = O(1)$, $\sup_{\eta\in\mathcal{B}} \left\|\Omega_\eta^{-1}\right\|_\eta = O(1)$, $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} \|\nabla_\theta \log f_{\theta_0}(Y)\|_\eta^{2+\nu} = O(1)$, *for some* $\nu > 0$.

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_{θ_0} are asymptotically of the same order, so that MSE provides a meaningful balance between bias and variance also asymptotically. Part (ii) requires δ_{θ_0} to be sufficiently smooth in θ_0, so that a first-order Taylor expansion provides a good local approximation of δ_{θ_0}. Part (iii) imposes similar smoothness assumption on $f_{\theta_0}(y)$ in θ_0. The first condition in part (iii) is just continuity in Hellinger distance, and the second condition is very similar, but also involves the score of the model. The last condition in part (iii) is a standard condition of differentiability in quadratic mean (see, e.g., equation (5.38) in Van der Vaart, 2000). Part (iv) of the assumption requires that our distance measure $d(\theta, \theta(\eta))$ converges to the associated norm for small values ϵ in a smooth fashion. Finally, part (v) requires invertibility of H_η^{-1} and Ω_η^{-1}, and uniform boundedness of various derivatives and of the $(2+\nu)$-th moment of $\nabla_\theta \log f_{\theta(\eta)}(y)$. Notice that invertibility of $H_{\theta(\eta)}$ is *not* required for our results.

For many of the proofs (specifically, all results below up to Proposition A1) we only need the regularity conditions in Assumption A1. However, in order to describe the properties of our Minimum-MSE estimator $\widehat{\delta}_\epsilon^{\text{MMSE}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\text{MMSE}}(Y_i, \widehat{\eta})$ we also need to account for the fact that $\widehat{\eta}$ is itself already an estimator. It turns our that the leading-order asymptotic properties of $\widehat{\delta}_\epsilon^{\text{MMSE}}$ are actually independent of whether η is known or estimated in the construction of $\widehat{\delta}_\epsilon^{\text{MMSE}}$ (see Lemma A3 below), but formally showing this requires some additional assumptions, which we present next.

For a given η, let $\mathcal{H}(\eta)$ be the set of functions $h = h(\cdot, \eta)$ that satisfy the constraints (2) and (4). The minimization problem (12) in the main text can then be rewritten as

$$\begin{aligned} Q_\epsilon^{\text{MMSE}}(\eta) &:= \min_{h\in\mathcal{H}(\eta)} \left[b_\epsilon(h,\eta)^2 + \frac{\text{Var}_{\theta(\eta)}(h(Y,\eta))}{n} \right] \\ &= b_\epsilon(h_\epsilon^{\text{MMSE}},\eta)^2 + \frac{\text{Var}_{\theta(\eta)}(h_\epsilon^{\text{MMSE}}(Y,\eta))}{n}. \end{aligned} \tag{A1}$$

The optimal $h_\epsilon^{\text{MMSE}}(\cdot,\eta) \in \mathcal{H}(\eta)$ can be expressed as

$$h_\epsilon^{\text{MMSE}}(y,\eta) = \langle v_\epsilon^{\text{MMSE}}(\eta), \nabla_\theta \log f_{\theta(\eta)}(y)\rangle, \tag{A2}$$

with

$$v_\epsilon^{\text{MMSE}}(\eta) := G_\eta' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + \left[\mathbb{I} - G_\eta' H_\eta^{-1} G_\eta H_{\theta(\eta)}\right] \left[\widetilde{H}_{\theta(\eta)} + (\epsilon n)^{-1} \Omega_\eta\right]^{-1} \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \quad \text{(A3)}$$

where $v_\epsilon^{\text{MMSE}}(\eta) \in T(\Theta)$, and $\mathbb{I}$ denotes the identity operator on $T(\Theta)$. It is easy to verify that $h_\epsilon^{\text{MMSE}}(y, \eta)$ in (A2) indeed satisfies the first-order conditions of problem (12).

Assumption A2. *We assume that*

(i) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left(\mathbb{E}_{\theta_0} \|\widehat{\eta} - \eta\|^4\right)^{1/4} = o(n^{-1/4})$.

(ii) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} \left\|\nabla_\eta h_\epsilon^{\text{MMSE}}(Y, \eta)\right\|^2 = O(1)$.

(iii) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} \sup_{\tilde{\eta}\in B(\eta,r_\epsilon)} \left\|\frac{1}{n}\sum_{i=1}^n \nabla_{\eta\eta'} h_\epsilon^{\text{MMSE}}(Y_i, \tilde{\eta})\right\|^2 = O(1)$, *for a Euclidean ball* $B(\eta, r_\epsilon)$ *around* η *with radius* $r_\epsilon = o(1)$.

Part (i) of Assumption A2 requires $\widehat{\eta}$ to converge at a rate faster than $n^{1/4}$, although in most applications we actually expect it to converge at rate $n^{1/2}$.[20] Part (ii) of Assumption A2 requires a uniformly bounded second moment for $\nabla_\eta h_\epsilon^{\text{MMSE}}(y, \eta)$. Since (A2) and (A3) give an explicit expression for $h_\epsilon^{\text{MMSE}}(y, \eta)$, we could replace Assumption A2(ii) by appropriate assumptions on the model primitives $f_{\theta_0}(y)$, δ_{θ_0} and Ω_η, but for the sake of brevity we state the assumption in terms of $h_\epsilon^{\text{MMSE}}(y, \eta)$. The same is true for part (iii) of Assumption A2. Notice that this last part of the assumption involves a supremum over $\tilde{\eta}$ inside of an expectation – in order to verify it, one either requires a uniform Lipschitz bound on the dependence of $h_\epsilon^{\text{MMSE}}(Y_i, \eta)$ on η, or some empirical process method to control the entropy of that function (e.g., a bracketing argument). But since η is a finite-dimensional parameter these are all standard arguments.

Remark. We found that $h_\epsilon^{\text{MMSE}}(y, \eta)$ can be expressed in the form $\langle v(\eta), \nabla_\theta \log f_{\theta(\eta)}(y)\rangle$, thus automatically satisfying the constraint (2). By choosing $v(\eta) = G_\eta' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + \widetilde{v}(\eta)$, where $G_\eta H_{\theta(\eta)} \widetilde{v}(\eta) = 0$, the constraint (4) is also satisfied. Using this one can alternatively represent the worst-case MSE problem as

$$Q_\epsilon^{\text{MMSE}}(\eta) = \min_{\widetilde{v}\in T(\Theta)} \left[\epsilon \left\|\widetilde{\nabla}_\theta \delta_{\theta(\eta)} - \widetilde{H}_{\theta(\eta)} \widetilde{v}\right\|_\eta^2 + \frac{1}{n} \left\langle \widetilde{v}, \widetilde{H}_{\theta(\eta)} \widetilde{v} \right\rangle\right] + \frac{1}{n} \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)}.$$

[20] By slightly modifying the proof of Lemma A3 below one could replace Assumption A2(i) by $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left(\mathbb{E}_{\theta_0} \|\widehat{\eta} - \eta\|^2\right)^{1/2} = o(n^{-1/2})$ – i.e., convergence in ℓ^2 only, but at a faster rate – although this would require slightly different versions of parts (ii) and (iii) of that assumption as well.

This concise expression for the leading order worst-case MSE highlights the terms of order ϵ (from squared bias) and of order $1/n$ (from variance terms). This representation also shows that instead of solving for the optimal influence function $h(y,\eta)$ we can alternatively solve for an optimal vector $\widetilde{v} \in T(\Theta)$, which is particularly convenient in models where the dimension of y exceeds that of θ.

A.1.2 Proof

In the following, as in the rest of the paper, we always implicitly assume that all functions of y are measurable, and that correspondingly all expectations and integrals over y are well-defined.

Lemma A1. *Let Assumption A1 and the conditions on $h_\epsilon(\cdot,\eta)$ in Theorem 1 hold. Then,*

(i) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left|\mathbb{E}_{\theta_0} h_\epsilon^2(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon^2(Y,\eta)\right| = o(1).$

(ii) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left|\mathbb{E}_{\theta_0} h_\epsilon(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) - \left\langle \theta_0 - \theta(\eta), \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) \right\rangle\right| = o(\epsilon^{1/2}).$

Proof of Lemma A1. # Part (i): Without loss of generality we may assume that $\kappa \leq 4$, since if $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} |h_\epsilon(Y,\eta)|^\kappa = O(1)$ holds for $\kappa > 4$, then it also holds for $\kappa \leq 4$. Let $\xi = \kappa/(\kappa-2) \geq 2$. We then have

$$\int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^\xi dy \leq \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right]^2 dy,$$

where we used that $|a-b| \leq |a^c - b^c|^{1/c}$, for any $a, b \geq 0$ and $c \geq 1$, and plugged in $a = f_{\theta_0}^{1/\xi}(y)$, $b = f_{\theta(\eta)}^{1/\xi}(y)$, and $c = \xi/2$. Thus, the first part of Assumption A1(iii) also implies

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^\xi dy = o(1). \tag{A4}$$

Next, we find

$$
\begin{aligned}
&\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left|\mathbb{E}_{\theta_0} h_\epsilon^2(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon^2(Y,\eta)\right| \\
&= \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| \int_{\mathcal{Y}} h_\epsilon^2(y,\eta) \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y)} \left[f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right] dy \right| \\
&\leq \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} |h_\epsilon(y,\eta)|^{\frac{2\xi}{\xi-1}} \left| \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y)} \right|^{\frac{\xi}{\xi-1}} dy \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\leq \xi \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} |h_\epsilon(y,\eta)|^{\frac{2\xi}{\xi-1}} \left| f_{\theta_0}(y) + f_{\theta(\eta)}(y) \right| dy \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\leq \xi \left\{ 2 \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} |h_\epsilon(Y,\eta)|^{\kappa} \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} = o(1),
\end{aligned}
$$

where the first inequality is an application of Hölder's inequality, the second inequality uses that $\left| \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y)} \right|^{\xi/(\xi-1)} \leq \xi^{\xi/(\xi-1)} \left[f_{\theta_0}(y) + f_{\theta(\eta)}(y) \right]$,[21] the last line uses that $\kappa = 2\xi/(\xi-1)$, and the final conclusion follows from our assumptions and (A4).

Part (ii): We have

$$
\begin{aligned}
&\mathbb{E}_{\theta_0} h_\epsilon(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) - \left\langle \theta_0 - \theta(\eta), \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) \right\rangle \\
&= \int_{\mathcal{Y}} h_\epsilon(y,\eta) \left[f_{\theta_0}(y) - f_{\theta(\eta)}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y) \right\rangle f_{\theta(\eta)}(y) \right] dy \\
&= \underbrace{\int_{\mathcal{Y}} h_\epsilon(y,\eta) \left[f_{\theta_0}^{1/2}(y) + f_{\theta(\eta)}^{1/2}(y) \right] \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \frac{1}{2} \left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y) \right\rangle f_{\theta(\eta)}^{1/2}(y) \right] dy}_{=: a_{\eta,\theta_0,q}^{(1)}} \\
&\quad + \frac{1}{2} \underbrace{\int_{\mathcal{Y}} h_\epsilon(y,\eta) f_{\theta(\eta)}^{1/2}(y) \left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y) \right\rangle \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right] dy}_{=: a_{\eta,\theta_0,q}^{(2)}}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| a_{\eta,\theta_0,q}^{(1)} \right|^2 \\
&\leq 4 \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} h_\epsilon^2(Y,\eta) \right\} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y) \right\rangle \right]^2 dy \right\} \\
&= o(\epsilon),
\end{aligned}
$$

[21] For $a, b \geq 0$ there exists $c \in [a,b]$ such that by the mean value theorem we have $(a^\xi - b^\xi)/(a-b) = \xi c^{\xi-1} \leq \xi \max(a^{\xi-1}, b^{\xi-1})$, and therefore $[(a^\xi - b^\xi)/(a-b)]^{\xi/(\xi-1)} \leq \xi^{\xi/(\xi-1)} \max(a^\xi, b^\xi) \leq \xi^{\xi/(\xi-1)}(a^\xi + b^\xi)$, which we apply here with $a = f_{\theta_0}^{1/\xi}(y)$ and $b = f_{\theta(\eta)}^{1/\xi}(y)$.

and

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\left|a^{(2)}_{\eta,\theta_0,q}\right|^2$$
$$\leq \left\{\mathbb{E}_{\theta(\eta)}h^2_\epsilon(Y,\eta)\right\}\left\{\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\|\theta_0-\theta(\eta)\|^2_{\text{ind},\eta}\int_{\mathcal{Y}}\left\|\nabla_\theta \log f_{\theta(\eta)}(y)\right\|^2_\eta\left[f^{1/2}_{\theta_0}(y)-f^{1/2}_{\theta(\eta)}(y)\right]^2 dy\right\}$$
$$= o(\epsilon).$$

Combining this gives the statement in the lemma. ∎

Let $\Delta_{\eta,\theta_0} := \delta_{\theta_0} - \delta_{\theta(\eta)}$. For a function $h = h(y,\eta)$ we define

$$\begin{aligned} Q_\epsilon(h,\eta,\theta_0) &:= \mathbb{E}_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^n h(Y_i,\eta) - \Delta_{\eta,\theta_0}\right)^2 \\ &= \left[\mathbb{E}_{\theta_0}h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \frac{1}{n}\text{Var}_{\theta_0}\left[h(Y,\eta) - \Delta_{\eta,\theta_0}\right] \\ &= \frac{n-1}{n}\left[\mathbb{E}_{\theta_0}h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \frac{1}{n}\mathbb{E}_{\theta_0}\left[h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2. \end{aligned} \tag{A5}$$

Also, recall the definition of the worst-case bias in (8) of the main text:

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

Lemma A2. *Let Assumption A1 and the conditions on $h_\epsilon(\cdot,\eta)$ in Theorem 1 hold. Then,*

$$\sup_{\eta\in\mathcal{B}}\left|\sup_{\theta_0\in\Gamma_\epsilon(\eta)} Q_\epsilon(h_\epsilon,\eta,\theta_0) - b_\epsilon(h_\epsilon,\eta)^2 - \frac{\text{Var}_{\theta(\eta)}(h_\epsilon(Y,\eta))}{n}\right| = o(\epsilon).$$

Proof of Lemma A2. Using the Cauchy-Schwarz inequality and our assumptions we find that

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\left|\left\langle\theta_0-\theta(\eta), \mathbb{E}_{\theta(\eta)}h_\epsilon(Y,\eta)\nabla_\theta \log f_{\theta(\eta)}(Y)\right\rangle\right|$$
$$\leq \sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\left\{\|\theta_0-\theta(\eta)\|_{\text{ind},\eta}\left[\mathbb{E}_{\theta(\eta)}h^2_\epsilon(Y,\eta)\right]^{1/2}\left[\mathbb{E}_{\theta(\eta)}\left\|\nabla_\theta \log f_{\theta(\eta)}(Y)\right\|^2_\eta\right]^{1/2}\right\} = o(1), \tag{A6}$$

and similarly

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\left|\left\langle\theta_0-\theta(\eta), \nabla_\theta\delta_{\theta(\eta)}\right\rangle\right| = o(1). \tag{A7}$$

Lemma A1(ii) and (A6) imply that $\mathbb{E}_{\theta_0}h_\epsilon(Y,\eta) = \mathbb{E}_{\theta(\eta)}h_\epsilon(Y,\eta) + o(1)$, uniformly in $(\theta_0,\eta)\in\Gamma_\epsilon$. In turn, Assumption A1(ii) guarantees that $\Delta_{\eta,\theta_0} = o(1)$, uniformly in $(\theta_0,\eta)\in\Gamma_\epsilon$. Combining with Lemma A1(i) we thus obtain

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[h_\epsilon(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 &= \mathbb{E}_{\theta_0}\left[h_\epsilon(Y,\eta)\right]^2 - 2\,\Delta_{\eta,\theta_0}\,\mathbb{E}_{\theta_0}h_\epsilon(Y,\eta) + \Delta^2_{\eta,\theta_0} \\ &= \mathbb{E}_{\theta(\eta)}\left[h_\epsilon(Y,\eta)\right]^2 + o(1) = \text{Var}_{\theta(\eta)}(h_\epsilon(Y,\eta)) + o(1),\end{aligned}$$

uniformly in $(\theta_0, \eta) \in \Gamma_\epsilon$, where in the last step we have also used that $h_\epsilon(y, \eta)$ satisfies the unbiasedness constraint (2). Using that constraint again, as well as Lemma A1(ii) and Assumptions A1(ii) and A1(iv) we find

$$\begin{aligned}
&\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} |\mathbb{E}_{\theta_0} h_\epsilon(Y, \eta) - \Delta_{\eta,\theta_0}| \\
&\quad = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \left| \left\langle \theta_0 - \theta(\eta), \mathbb{E}_{\theta(\eta)} h_\epsilon(Y, \eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)} \right\rangle \right| + o(\epsilon^{1/2}) \\
&\quad = \epsilon^{1/2} \left\| \mathbb{E}_{\theta(\eta)} h_\epsilon(Y, \eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)} \right\|_\eta + o(\epsilon^{1/2}) = b_\epsilon(h_\epsilon, \eta) + o(\epsilon^{1/2}),
\end{aligned}$$

uniformly in $\eta \in \mathcal{B}$. The results in the previous two displays together with the last expression for $Q_\epsilon(h, \eta, \theta_0)$ in equation (A5) yield the statement of the lemma. ∎

Proposition A1. *Let Assumption A1 and the conditions on $h_\epsilon(\cdot, \eta)$ in Theorem 1 hold. Then,*

$$\sup_{\eta \in \mathcal{B}} \left\{ Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0} \right)^2 \right] \right\} \leq o(\epsilon).$$

Proof of Proposition A1. Using (20), the definition of $Q_\epsilon(h, \eta, \theta_0)$, and also applying Lemma A2, we find that

$$\begin{aligned}
&\sup_{\eta \in \mathcal{B}} \left\{ Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0} \right)^2 \right] \right\} \\
&= \sup_{\eta \in \mathcal{B}} \left\{ Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\frac{1}{n} \sum_{i=1}^n h_\epsilon(Y_i, \eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} \right)^2 \right] \right\} + o(\epsilon) \\
&= \sup_{\eta \in \mathcal{B}} \left[Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} Q_\epsilon(h_\epsilon, \eta, \theta_0) \right] + o(\epsilon) \\
&= \sup_{\eta \in \mathcal{B}} \left[Q_\epsilon^{\text{MMSE}}(\eta) - b_\epsilon(h_\epsilon, \eta)^2 - \frac{\text{Var}_{\theta(\eta)}(h_\epsilon(Y, \eta))}{n} \right] + o(\epsilon).
\end{aligned}$$

Moreover, by the definition of $Q_\epsilon^{\text{MMSE}}(\eta)$ in (A1) we have

$$Q_\epsilon^{\text{MMSE}}(\eta) \leq b_\epsilon(h_\epsilon, \eta)^2 + \frac{\text{Var}_{\theta(\eta)}(h_\epsilon(Y, \eta))}{n}.$$

Combining the last two displays gives the statement of the proposition. ∎

Recall that $\widehat{\delta}_\epsilon^{\,\text{MMSE}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n} \sum_{i=1}^n h_\epsilon^{\text{MMSE}}(Y_i, \widehat{\eta})$. The following lemma shows that the fact that η is being estimated in the construction of $\widehat{\delta}_\epsilon^{\,\text{MMSE}}$ can be neglected to first order. Notice that this result requires the additional regularity conditions in Assumption A2, which were not required for any of the previous results.

Lemma A3. *Under Assumptions A1 and A2 we have*

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\left[\widehat{\delta}_\epsilon^{\,\mathrm{MMSE}} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta)\right]^2 = o(\epsilon).$$

Proof of Lemma A3. By a Taylor expansion in η we find that

$$\begin{aligned}
\widehat{\delta}_\epsilon^{\,\mathrm{MMSE}} &= \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\widehat{\eta}) \\
&= \delta_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta) + \underbrace{(\widehat{\eta}-\eta)'\left[\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right]}_{=r^{(1)}_{\eta,\theta_0}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta)'\frac{1}{n}\sum_{i=1}^n\left[\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta) - \mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta)\right]}_{=r^{(2)}_{\eta,\theta_0}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta)'\left[\mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) - \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right]}_{=r^{(3)}_{\eta,\theta_0}} \\
&\quad + \frac{1}{2}\underbrace{(\widehat{\eta}-\eta)'\left[\frac{1}{n}\sum_{i=1}^n \nabla_{\eta\eta'} h_\epsilon^{\mathrm{MMSE}}(Y_i,\widetilde{\eta})\right]}_{=r^{(4)}_{\eta,\theta_0}}(\widehat{\eta}-\eta),
\end{aligned} \tag{A8}$$

where $\widetilde{\eta}$ is a value between $\widehat{\eta}$ and η. Our constraints (2) and (4) guarantee that $\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) = 0$, that is, we have $r^{(1)}_{\eta,\theta_0} = 0$. Using Assumption A2 we furthermore find

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left|r^{(2)}_{\eta,\theta_0}\right|^2 &\le \mathbb{E}_{\theta_0}\|\widehat{\eta}-\eta\|^2\ \mathbb{E}_{\theta_0}\left\|\frac{1}{n}\sum_{i=1}^n\left[\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta) - \mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta)\right]\right\|^2 \\
&\le \mathbb{E}_{\theta_0}\|\widehat{\eta}-\eta\|^2\ \frac{1}{n}\,\mathbb{E}_{\theta_0}\left\|\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right\|^2 = o(n^{-1/2})O(n^{-1}) = o(\epsilon^{3/2}) = o(\epsilon),
\end{aligned}$$

uniformly in $(\theta_0,\eta)\in\Gamma_\epsilon$, where in the second step we have used the independence of Y_i across i. Similarly, we have

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left|r^{(3)}_{\eta,\theta_0}\right|^2 &\le \mathbb{E}_{\theta_0}\|\widehat{\eta}-\eta\|^2\left\|\mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) - \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right\|^2 \\
&= o(n^{-1/2})O(\epsilon) = o(\epsilon^{3/2}) = o(\epsilon),
\end{aligned}$$

uniformly in $(\theta_0,\eta)\in\Gamma_\epsilon$, where we have used that

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\left\|\mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) - \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right\| = O(\epsilon^{1/2}),$$

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

$$\mathbb{E}_{\theta_0}\left|r^{(4)}_{\eta,\theta_0}\right|^2 \leq \mathbb{E}_{\theta_0}\left\|\widehat{\eta}-\eta\right\|^4 \mathbb{E}_{\theta_0}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{\eta\eta'}h^{\text{MMSE}}_{\epsilon}(Y_i,\widetilde{\eta})\right\|^2 = o(n^{-1}) = o(\epsilon),$$

uniformly in $(\theta_0,\eta)\in\Gamma_\epsilon$, where we have used Assumption A2(iii). We have thus shown that

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\left|r^{(1)}_{\eta,\theta_0}+r^{(2)}_{\eta,\theta_0}+r^{(3)}_{\eta,\theta_0}+\frac{1}{2}r^{(4)}_{\eta,\theta_0}\right|^2 = o(\epsilon),$$

which together with (A8) gives the statement of the lemma. ∎

Proposition A2. *Under Assumptions A1 and A2 we have*

$$\sup_{\eta\in\mathcal{B}}\left|\sup_{\theta_0\in\Gamma_\epsilon(\eta)}\mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}^{\text{MMSE}}_{\epsilon}-\delta_{\theta_0}\right)^2\right]-Q^{\text{MMSE}}_{\epsilon}(\eta)\right| = o(\epsilon).$$

Proof of Proposition A2. Applying Lemma A3 together with the definition of $Q_\epsilon(h,\eta,\theta_0)$ in (A5) we obtain

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

Assumptions A1(i) and A1(v) imply that $\sup_{\eta\in\mathcal{B}}\left\|v^{\text{MMSE}}_{\epsilon}(\eta)\right\|_{\text{ind},\eta} = O(1)$.[22] From the explicit solution for $h^{\text{MMSE}}_{\epsilon}(y,\eta)$ in (A2) and (A3) together with Assumption A2 we conclude

$$\begin{aligned}\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\left[h^{\text{MMSE}}_{\epsilon}(Y,\eta)\right]^{2+\nu} &= \sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\langle v^{\text{MMSE}}_{\epsilon}(\eta),\nabla_\theta\log f_{\theta(\eta)}(y)\rangle^{2+\nu}\\ &\leq \sup_{\eta\in\mathcal{B}}\left\|v^{\text{MMSE}}_{\epsilon}(\eta)\right\|^{2+\nu}_{\text{ind},\eta}\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\left\|\nabla_\theta\log f_{\theta_0}(Y)\right\|^{2+\nu}_{\eta} = O(1).\end{aligned}$$

Thus, $h^{\text{MMSE}}_{\epsilon}(y,\eta)$ satisfies the regularity conditions for $h_\epsilon(y,\eta)$ in Theorem 1 with $\kappa = 2+\nu$. We can therefore apply Lemma A2 with $h_\epsilon(y,\eta) = h^{\text{MMSE}}_{\epsilon}(y,\eta)$ to find

$$\sup_{\eta\in\mathcal{B}}\left|\sup_{\theta_0\in\Gamma_\epsilon(\eta)}Q_\epsilon(h^{\text{MMSE}}_{\epsilon},\eta,\theta_0)-Q^{\text{MMSE}}_{\epsilon}(\eta)\right| = o(\epsilon).$$

Combining the last two displays gives the statement of the proposition. ∎

Proof of Theorem 1. Combining Propositions A1 and A2 gives the the statement of the theorem. ∎

[22] Notice that $\sup_{\eta\in\mathcal{B}}\left\|H_{\theta(\eta)}\right\|_\eta = O(1)$ follows from the bounded moment condition on the score $\nabla_\theta\log f_{\theta(\eta)}(y)$ in part (v) of Assumption A1.

A.2 Proof of Corollary 1

Let $q(\eta)$ denote the MSE difference in the curly brackets in (21). Corollary 1 then immediately follows from Theorem 1 and $\int_{\mathcal{B}} q(\eta)w(\eta)d\eta \leq \left[\int_{\mathcal{B}} w(\eta)d\eta\right]\left[\sup_{\eta\in\mathcal{B}} q(\eta)\right]$.

A.3 Proof of Theorem 2

Assumption A3.

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

(ii) $\widehat{\delta} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta) = o_{P_{\theta_0}}(n^{-\frac{1}{2}})$, *uniformly in* $(\theta_0, \eta) \in \Gamma_\epsilon$.

(iii) *Let* $\sigma_h^2(\theta_0, \eta) = \mathrm{Var}_{\theta_0} h(Y, \eta)$. *We assume that there exists a constant* c, *independent of* ϵ, *such that* $\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \sigma_h(\theta_0, \eta) \geq c > 0$. *Furthermore, for all sequences* $a_n = c_{1-\mu/2} + o(1)$ *we have*

$$\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \Pr_{\theta_0}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{h(Y_i, \eta) - \mathbb{E}_{\theta_0} h(Y, \eta)}{\sigma_h(\theta_0, \eta)}\right| \leq a_n\right] \geq 1 - \mu + o(1).$$

(iv) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\|\widehat{\eta} - \eta\|^2 = o(1)$, $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}[\widehat{\sigma}_h - \sigma_h(\theta_0, \eta)]^2 = o(1)$.

(v) $\sup_{\eta\in\mathcal{B}} \|\nabla_\eta b_\epsilon(h, \eta)\| = O(\epsilon^{\frac{1}{2}})$.

Part (ii) is weaker than the local regularity of the estimator $\widehat{\delta}$ that we assumed when analyzing the minimum-MSE estimator, see equation (20). In turn, related to but differently from the conditions we used for Theorem 1, part (iii) requires a form of local asymptotic normality of the estimator.

Proof of Theorem 2. Let $\widehat{\delta}$ be an estimator and $h(y, \eta)$ be the corresponding influence function such that part (ii) in Assumption A3 holds. Define $\widehat{R}_\eta := \widehat{\delta} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta)$. We then have

$$\begin{aligned}\widehat{\delta} - \delta_{\theta_0} &= \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} + \widehat{R}_\eta \\ &= \frac{1}{n}\sum_{i=1}^{n} [h(Y_i, \eta) - \mathbb{E}_{\theta_0} h(Y, \eta)] - \left[\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y, \eta)\right] + \widehat{R}_\eta,\end{aligned}$$

and therefore

$$\underbrace{\frac{|\widehat{\delta} - \delta_{\theta_0}| - b_\epsilon(h, \widehat{\eta}) - \widehat{\sigma}_h\, c_{1-\mu/2}/\sqrt{n}}{\sigma_h(\theta_0, \eta)/\sqrt{n}}}_{=\text{lhs}} \leq \underbrace{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{h(Y_i, \eta) - \mathbb{E}_{\theta_0} h(Y, \eta)}{\sigma_h(\theta_0, \eta)}\right| - c_{1-\mu/2} + \widehat{r}_{\eta,\theta_0}}_{=\text{rhs}}, \quad \text{(A9)}$$

where

$$\widehat{r}_{\eta,\theta_0} := c_{1-\mu/2} + \frac{\left|\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)\right| + \left|\widehat{R}_\eta\right| - b_\epsilon(h,\widehat{\eta}) - \widehat{\sigma}_h\, c_{1-\mu/2}/\sqrt{n}}{\sigma_h(\theta_0,\eta)/\sqrt{n}}$$
$$= \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \left\{ |\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)| + |\widehat{R}_\eta| - b_\epsilon(h,\widehat{\eta}) - \frac{\widehat{\sigma}_h - \sigma_h(\theta_0,\eta)}{\sqrt{n}} c_{1-\mu/2} \right\}.$$

From (A9) we conclude that the event rhs ≤ 0 implies the event lhs ≤ 0, and therefore $\Pr_{\theta_0}(\text{lhs} \leq 0) \geq \Pr_{\theta_0}(\text{rhs} \leq 0)$, which we can also write as

$$\Pr_{\theta_0}\left[|\widehat{\delta} - \delta_{\theta_0}| \leq b_\epsilon(h,\widehat{\eta}) + \frac{\widehat{\sigma}_h}{\sqrt{n}} c_{1-\mu/2} \right]$$
$$\geq \Pr_{\theta_0}\left[\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{h(Y_i,\eta) - \mathbb{E}_{\theta_0} h(Y,\eta)}{\sigma_h(\theta_0,\eta)} \right| \leq c_{1-\mu/2} - \widehat{r}_{\eta,\theta_0} \right]. \tag{A10}$$

By part (v) in Assumption A3 there exists a constant $C > 0$ such that $\sup_{\eta\in\mathcal{B}} \|\nabla_\eta b_\epsilon(h,\eta)\| \leq C\epsilon^{\frac{1}{2}}$, and therefore

$$\sup_{\eta\in\mathcal{B}} |b_\epsilon(h,\widehat{\eta}) - b_\epsilon(h,\eta)| \leq C\,\epsilon^{\frac{1}{2}}\, \|\widehat{\eta} - \eta\|.$$

Using this we find that

$$|\widehat{r}_{\eta,\theta_0}| \leq \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \left\{ \left| |\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)| - b_\epsilon(h,\eta) \right| \right.$$
$$\left. + \frac{|\widehat{\sigma}_h - \sigma_h(\theta_0,\eta)|}{\sqrt{n}} c_{1-\mu/2} + C\,\epsilon^{\frac{1}{2}}\, \|\widehat{\eta} - \eta\| + |\widehat{R}_\eta| \right\}.$$

Parts (ii) and (iii) in Assumption A3 imply that, uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$, we have

$$\frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \widehat{R}_\eta = o_{P_{\theta_0}}(1),$$

and analogously we find from the conditions in Assumption A3 that

$$\frac{\widehat{\sigma}_h - \sigma_h(\theta_0,\eta)}{\sigma_h(\theta_0,\eta)} = o_{P_{\theta_0}}(1), \qquad \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \epsilon^{\frac{1}{2}}\, \|\widehat{\eta} - \eta\| = o_{P_{\theta_0}}(1),$$

uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$. Finally, since we also impose Assumption A1 and $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} h^2(Y,\eta) = O(1)$ we obtain, analogously to the proof of Lemma A1(ii) above, that[23]

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \left| |\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)| - b_\epsilon(h,\eta) \right| = o(1).$$

We thus conclude that $\widehat{r}_{\eta,\theta_0} = o_{P_{\theta_0}}(1)$, uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$. Together with (A10) and part (iii) in Assumption A3 this implies (23), hence Theorem 2. ∎

[23] Notice that the proof of part (ii) of Lemma A1 only requires a bounded second moment of $h(y,\eta)$.

A.4 Asymptotically linear estimators

In this subsection we provide some background on the asymptotically linear representation (1), and we give several examples. See, e.g., Bickel *et al.* (1993) and Rieder (1994) on local asymptotic expansions of regular estimators.

Consider an asymptotically linear estimator $\widehat{\delta}$ which has the following representation under f_{θ_0}, for $\theta_0 \in \Theta$,

$$\widehat{\delta} = \delta^*_{\theta_0} + \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \theta_0) + o_{P_{\theta_0}}(n^{-\frac{1}{2}}), \tag{A11}$$

where $\delta^*_{\theta_0}$ is the probability limit of $\widehat{\delta}$ under f_{θ_0}, and $\phi(y, \theta_0)$ is its influence function. The pseudo-true value $\delta^*_{\theta_0}$ generally differs from the true parameter value δ_{θ_0}. The influence function is assumed to satisfy

$$\mathbb{E}_{\theta_0}\phi(Y, \theta_0) = 0, \quad \nabla_\theta \delta^*_{\theta_0} + \mathbb{E}_{\theta_0}\nabla_\theta \phi(Y, \theta_0) = 0, \quad \text{for all } \theta_0 \in \Theta. \tag{A12}$$

The first condition in (A12) requires that the estimator be asymptotically unbiased for the pseudo-true value $\delta^*_{\theta_0}$. The second condition is a version of the generalized information identity.[24] Expansion (A11) and conditions (A12) are satisfied for a large class of estimators, see below for examples.

Furthermore, suppose that

$$\delta^*_{\theta(\eta)} = \delta_{\theta(\eta)}, \quad \text{for all } \eta \in \mathcal{B}. \tag{A13}$$

Condition (A13) requires that $\widehat{\delta}$ be asymptotically unbiased for $\delta_{\theta(\eta)}$ under $f_{\theta(\eta)}$, that is, under correct specification of the reference model. Note that, under mild regularity conditions, the function

$$h(y, \eta) = \phi(y, \theta(\eta))$$

will then be automatically "locally robust" with respect to η, as defined in Chernozhukov *et al.* (2016). Indeed,

$$\begin{aligned}\mathbb{E}_{\theta(\eta)}\nabla_\eta h(Y, \eta) &= \mathbb{E}_{\theta(\eta)}\nabla_\eta \phi(y, \theta(\eta)) = \nabla_\eta \theta(\eta)\, \mathbb{E}_{\theta(\eta)}\nabla_\theta \phi(y, \theta(\eta)) \\ &= -\nabla_\eta \theta(\eta)\nabla_\theta \delta^*_{\theta(\eta)} = -\nabla_\eta \delta^*_{\theta(\eta)} = -\nabla_\eta \delta_{\theta(\eta)},\end{aligned}$$

where we have used (A12) at $\theta_0 = \theta(\eta)$, and that, by (A13), $\nabla_\eta \delta^*_{\theta(\eta)} = \nabla_\eta \delta_{\theta(\eta)}$.

[24] The generalized information identity can alternatively be written in terms of the influence function and the score of the model (or any parametric sub-model in semi-parametric settings); see, e.g., Newey (1990).

To relate (1), which is taken around $\delta_{\theta(\eta)}$, to expansion (A11), which is taken around $\delta^*_{\theta_0}$, note that by an expansion around $\theta(\eta)$, and making use of the second identity in (A12), (A11) will imply (1) provided $\frac{1}{n}\sum_{i=1}^n \nabla_\theta \phi(Y_i, \widetilde{\theta}) - \mathbb{E}_{\widetilde{\theta}} \nabla_\theta \phi(Y, \widetilde{\theta})$ is $o_{P_{\theta_0}}(1)$, uniformly in $\eta \in \mathcal{B}$, $\theta_0 \in \Gamma_\epsilon(\eta)$, $\widetilde{\theta} \in \Gamma_\epsilon(\eta)$.

Examples. As a first example, consider an estimator $\widehat{\delta}$ solving $\sum_{i=1}^n m(Y_i, \widehat{\delta}) = 0$, where m is a smooth scalar moment function. The pseudo-true value solves $\mathbb{E}_{\theta_0} m(Y, \delta^*_{\theta_0}) = 0$ for all $\theta_0 \in \Theta$. Expanding the moment condition around $\delta^*_{\theta_0}$ implies that (A11) holds under mild conditions on m, with

$$\phi(y, \theta_0) = \left[-\mathbb{E}_{\theta_0} \nabla_\delta m(Y, \delta^*_{\theta_0})\right]^{-1} m(y, \delta^*_{\theta_0}).$$

It is easy to see that (A12) is satisfied. Moreover, (A13) is satisfied when the moment restriction is satisfied under the reference model; that is, whenever $\mathbb{E}_{\theta(\eta)} m(Y, \delta_{\theta(\eta)}) = 0$ for all $\eta \in \mathcal{B}$.

As a second example, consider an estimator $\widehat{\delta}$ solving $\sum_{i=1}^n m(Y_i, \widehat{\delta}, \widehat{\eta}) = 0$, where $\widehat{\eta}$ is a preliminary estimator which solves $\sum_{i=1}^n q(Y_i, \widehat{\eta}) = 0$, for smooth moment functions m (scalar) and q (vector-valued). In this case (A11) holds under regularity conditions on m and q, with

$$\begin{aligned}\phi(y, \theta_0) =& [\mathbb{E}_{\theta_0}(-\nabla_\delta m(Y, \delta^*_{\theta_0}, \eta^*_{\theta_0}))]^{-1} \Big(m(y, \delta^*_{\theta_0}, \eta^*_{\theta_0}) \\ &+ \mathbb{E}_{\theta_0}(\nabla_\eta m(Y, \delta^*_{\theta_0}, \eta^*_{\theta_0}))' [\mathbb{E}_{\theta_0}(-\nabla_\eta q(Y, \eta^*_{\theta_0}))]^{-1} q(y, \eta^*_{\theta_0}) \Big),\end{aligned}$$

where $\eta^*_{\theta_0}$ and $\delta^*_{\theta_0}$ satisfy $\mathbb{E}_{\theta_0} q(Y, \eta^*_{\theta_0}) = 0$ and $\mathbb{E}_{\theta_0} m(Y, \delta^*_{\theta_0}, \eta^*_{\theta_0}) = 0$ for all $\theta_0 \in \Theta$. It can be verified that (A12) holds. Moreover, (A13) holds provided the moment restrictions for η and $\delta_{\theta(\eta)}$ are satisfied under the reference model, that is, whenever $\mathbb{E}_{\theta(\eta)} q(Y, \eta) = 0$ and $\mathbb{E}_{\theta(\eta)} m(Y, \delta_{\theta(\eta)}, \eta) = 0$ for all $\eta \in \mathcal{B}$.

As a third example, consider the (non-random) estimator $\widehat{\delta} = \delta_{\theta(\eta)}$, where η is a known, fixed parameter (i.e., $\mathcal{B} = \{\eta\}$). In this case $\phi(y, \theta_0) = \delta_{\theta(\eta)} - \delta^*_{\theta_0} = 0$. It follows that both (A12) and (A13) hold.

As a last example, consider the estimator $\widehat{\delta} = \delta_{\theta(\widehat{\eta})}$, where as above the preliminary estimator $\widehat{\eta}$ solves $\sum_{i=1}^n q(Y_i, \widehat{\eta}) = 0$. In this case (A11) will hold, with

$$\phi(y, \theta_0) = (\nabla_\eta \delta_{\theta(\eta^*_{\theta_0})})' \left[\mathbb{E}_{\theta_0}(-\nabla_\eta q(Y, \eta^*_{\theta_0}))\right]^{-1} q(y, \eta^*_{\theta_0}),$$

where $\eta^*_{\theta_0}$ solves $\mathbb{E}_{\theta_0} q(Y, \eta^*_{\theta_0}) = 0$. It is easy to see that (A12) is satisfied. Moreover, (A13) holds provided $\mathbb{E}_{\theta(\eta)} q(Y, \eta) = 0$ for all $\eta \in \mathcal{B}$.

B Semi-parametric models

In this section of the appendix we provide results and additional examples for the semi-parametric setting of Section 4.

B.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 B4. *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 $\delta_+ > 0$.[25] 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 $k_*(t)$.

[25] 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 $t \in \mathbb{R}$.

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

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

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 $0 \leq \tilde{t} \leq t \leq \delta_+$. 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 $\bar{t} \in [0, t]$. ∎

B.2 Mapping between the setup of Section 2 and the semi-parametric case

In order to link the formulas in Section 2 to the ones we derived in Section 4 for semi-parametric models, let us focus for simplicity on the case where η is known and $\theta = \pi$ is the density of A. In this case, elements v of the tangent space satisfy $\int_{\mathcal{A}} v(a)da = 0$, and the corresponding squared norm is $\|v\|^2_{\text{ind},\eta} = \int_{\mathcal{A}} \frac{v(a)^2}{\pi_\gamma(a)} da$. Hence Ω_η is such that, for any two elements of the tangent space, $\langle w, \Omega_\eta v\rangle = \int_{\mathcal{A}} w(a) \frac{v(a)}{\pi_\gamma(a)} da$.

In turn, elements u of the co-tangent space satisfy $\int_{\mathcal{A}} u(a)\pi_\gamma(a)da = 0$. The squared dual norm is $\|u\|^2_\eta = \int_{\mathcal{A}} u^2(a)\pi_\gamma(a)da = \text{Var}_\gamma(u(A))$ (see Subsection B.1), and Ω_η^{-1} is such that $\langle \Omega_\eta^{-1} u, s\rangle = \int_{\mathcal{A}} u(a)s(a)\pi_\gamma(a)da = \text{Cov}_\gamma(u(A), s(A))$.

Next, $\nabla_\theta \delta_{\theta(\eta)}$ is an element of the co-tangent space such that, for all tangents v,

$$\langle v, \nabla_\theta \delta_{\theta(\eta)}\rangle = \int_{\mathcal{A}} v(a)(\Delta(a) - \delta_{\theta(\eta)})da.$$

We identify $\nabla_\theta \delta_{\theta(\eta)}$ with $\Delta - \delta_{\theta(\eta)}$. In turn, $\nabla_\theta \log f_{\theta(\eta)}(y)$ is an element of the co-tangent space such that, for all tangents v,

$$\langle v, \nabla_\theta \log f_{\theta(\eta)}(y)\rangle = \frac{\int_{\mathcal{A}} g_\beta(y \,|\, a)v(a)da}{\int_{\mathcal{A}} g_\beta(y \,|\, a)\pi_\gamma(a)da} - \int_{\mathcal{A}} v(a)da = \mathbb{E}_{\beta,\gamma}\left(\frac{v(A)}{\pi_\gamma(A)} \,|\, y\right) - \mathbb{E}_\gamma\left(\frac{v(A)}{\pi_\gamma(A)}\right).$$

We identify $\nabla_\theta \log f_{\theta(\eta)}(y)$ with $\frac{g_\beta(y \,|\, \cdot)}{\int_{\mathcal{A}} g_\beta(y \,|\, a)\pi_\gamma(a)da} - 1$.

For any tangent v, $H_{\theta(\eta)} v$ is a co-tangent element such that, for all tangents w,

$$\begin{aligned}\langle w, H_{\theta(\eta)} v\rangle =& \mathbb{E}_{\theta(\eta)} \langle v, \nabla_\theta \log f_{\theta(\eta)}(Y)\rangle \langle w, \nabla_\theta \log f_{\theta(\eta)}(Y)\rangle \\ =& \text{Cov}_{\beta,\gamma}\left[\mathbb{E}_{\beta,\gamma}\left(\frac{v(A)}{\pi_\gamma(A)} \,|\, Y\right), \mathbb{E}_{\beta,\gamma}\left(\frac{w(A)}{\pi_\gamma(A)} \,|\, Y\right)\right].\end{aligned}$$

In particular, it follows that defining h_ϵ^{MMSE} as in (29) gives the same expression as in (41) since, for all y,

$$\begin{aligned} h_\epsilon^{\text{MMSE}}(y) &= \langle \left[H_{\theta(\eta)} + (\epsilon n)^{-1}\Omega_\eta\right]^{-1} \nabla_\theta \delta_{\theta(\eta)}, \nabla_\theta \log f_{\theta(\eta)}(y)\rangle \\ &= \mathbb{E}_{\beta,\gamma}\left[\frac{1}{\pi_\gamma(A)}\left[H_{\theta(\eta)} + (\epsilon n)^{-1}\Omega_\eta\right]^{-1} [\nabla_\theta \delta_{\theta(\eta)}](A) \,|\, y\right] \\ &= \mathbb{E}_{\beta,\gamma}\left[\left[\mathbb{H}_{\mathcal{A}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\right]^{-1} [\Delta - \delta](A) \,|\, y\right] \\ &= \mathbb{E}_{\mathcal{A}|\mathcal{Y}}\left[\mathbb{H}_{\mathcal{A}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\right]^{-1} [\Delta - \delta](y).\end{aligned}$$

We also briefly want to discuss when the conditions in Assumption A1(iii) are satisfied for this model. Firstly, we have

$$\int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y)\right]^2 dy = 2\, H^2(f_{\theta_0}, f_{\theta(\eta)}) \leq 2\, D_{\text{KL}}(f_{\theta_0} || f_{\theta(\eta)}) \leq 2\, D_{\text{KL}}(\pi_0 || \pi_\gamma),$$

where the first inequality is the general relation $H^2(f_{\theta_0}, f_{\theta(\eta)}) \leq D_{\mathrm{KL}}(f_{\theta_0}||f_{\theta(\eta)})$ between the squared Hellinger distance H^2 and the Kullback-Leibler divergence D_{KL}, and the second inequality is sometimes called the "chain rule" for the Kullback-Leibler divergence, which can be derived by an application of Jensen's inequality. Finally, recall that we defined our distance measure $d(\theta_0, \theta(\eta))$ in the semi-parametric case to be twice the Kullback-Leibler divergence $2D_{\mathrm{KL}}(\pi_0||\pi_\gamma) = 2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi_\gamma(a)}\right)\pi_0(a)da$. We therefore find that

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y)\right]^2 dy \leq \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} d(\theta_0, \theta(\eta))) = \epsilon = o(1),$$

that is, the first condition in Assumption A1(iii) is satisfied here. The second condition in that assumption follows if, for example, we assume that $\sup_{y\in\mathcal{Y}} \mathrm{Var}_\gamma\left[g_\beta(y \mid A)\right] / \left[\mathbb{E}_\gamma g_\beta(y \mid A)\right]^2 = O(1)$,[26] because an upper bound on $\left\|\nabla_\theta \log f_{\theta(\eta)}(y)\right\|_\eta^2 = \mathrm{Var}_\gamma\left[g_\beta(y \mid A)\right] / \left[\mathbb{E}_\gamma g_\beta(y \mid A)\right]^2$ can then simply be taken out of the integral over $y \in \mathcal{Y}$.

Regarding the last condition of Assumption A1(iii), we first note that since f_θ is linear in $\theta = \pi$ here we have $f_{\theta_0}(y) - f_{\theta(\eta)}(y) = \left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}(y)\right\rangle$, and therefore

$$\left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y)\right\rangle = \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{2f_{\theta(\eta)}^{1/2}(y)}.$$

Using this, we obtain

$$\begin{aligned}
&\int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y)\right\rangle\right]^2 dy \\
&= \frac{1}{4}\int_{\mathcal{Y}} \left[\left(\frac{f_{\theta_0}(y)}{f_{\theta(\eta)}(y)}\right)^{1/2} - 1\right]^4 f_{\theta(\eta)}(y)dy \leq \frac{1}{64}\int_{\mathcal{Y}} \left[\frac{f_{\theta_0}(y)}{f_{\theta(\eta)}(y)} - 1\right]^4 f_{\theta(\eta)}(y)dy \\
&= \frac{1}{64}\int_{\mathcal{Y}} \frac{\left[f_{\theta_0}(y) - f_{\theta(\eta)}(y)\right]^4}{\left[f_{\theta(\eta)}(y)\right]^3}dy = \frac{1}{64}\int_{\mathcal{Y}} \left[\left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y)\right\rangle\right]^4 f_{\theta(\eta)}(y)dy \\
&\leq \frac{1}{64}\left\|\theta_0 - \theta(\eta)\right\|_{\mathrm{ind},\eta}^4 \mathbb{E}_{\theta(\eta)}\left\|\nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta^4,
\end{aligned}$$

where the first inequality follows from $\sqrt{a} - 1 \leq (a-1)/2$ for $a \geq 0$. This shows that the last part of Assumption A1(iii) holds, provided $\mathbb{E}_{\theta(\eta)}\left\|\nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta^4 = O(1)$.

[26] If γ is not assumed known, then one should also take the supremum over γ in that condition.

B.3 Computation of $\epsilon(p)$

Given a probability p we set $\epsilon(p) = 4\Phi^{-1}(p)^2/(n\lambda_{max})$, where λ_{max} is the maximal eigenvalue of the projected Hessian operator which to $\pi(a')$ associates the function

$$\int_{\mathcal{Y}} \left[\left(\frac{g_\beta(y|a')}{\int_{\mathcal{A}} g_\beta(y|a)\pi_\gamma(a)da} - 1 \right) \left(\frac{\int_{\mathcal{A}} g_\beta(y|a)\pi(a)da}{\int_{\mathcal{A}} g_\beta(y|a)\pi_\gamma(a)da} - 1 \right) \int_{\mathcal{A}} g_\beta(y|a)\pi_\gamma(a)da \right] dy$$
$$- \left(\int_{\mathcal{Y}} \nabla_{\beta,\gamma} \log f_{\beta,\pi_\gamma}(y) g_\beta(y \,|\, a')dy \right)' H_\eta^{-1} \left(\int_{\mathcal{Y}} \nabla_{\beta,\gamma} \log f_{\beta,\pi_\gamma}(y) \int_{\mathcal{A}} g_\beta(y|a)\pi(a)dady \right).$$

That is, λ_{max} is the maximum of

$$\int_{\mathcal{Y}} \left[\left(\frac{\int_{\mathcal{A}} g_\beta(y|a)\pi(a)da}{\int_{\mathcal{A}} g_\beta(y|a)\pi_\gamma(a)da} - 1 \right)^2 \int_{\mathcal{A}} g_\beta(y|a)\pi_\gamma(a)da \right] dy$$
$$- \left(\int_{\mathcal{Y}} \nabla_{\beta,\gamma} \log f_{\beta,\pi_\gamma}(y) \int_{\mathcal{A}} g_\beta(y \,|\, a)\pi(a)dady \right)' H_\eta^{-1} \left(\int_{\mathcal{Y}} \nabla_{\beta,\gamma} \log f_{\beta,\pi_\gamma}(y) \int_{\mathcal{A}} g_\beta(y \,|\, a)\pi(a)dady \right),$$

with respect to π, subject to $\int_{\mathcal{A}} \pi(a)da = 1$ and $\int_{\mathcal{A}} \frac{(\pi(a)-\pi_\gamma(a))^2}{\pi_\gamma(a)}da \leq 1$.

Letting $\xi(a) = \pi(a)/\pi_\gamma(a)$, λ_{max} is thus equal to the maximum of

$$\mathbb{E}\left[\left(\mathbb{E}\left(\xi(A) \,|\, Y \right) - 1 \right)^2 \right] - \mathbb{E}\left[s_\eta(Y)\xi(A) \right]' H_\eta^{-1} \mathbb{E}\left[s_\eta(Y)\xi(A) \right]$$

with respect to ξ, subject to $\mathbb{E}\xi(A) = 1$ and $\operatorname{Var}\xi(A) \leq 1$, where we have denoted $s_\eta = \nabla_{\beta,\gamma} \log f_{\beta,\pi_\gamma}$, and in the remainder of this subsection we abstract from parameter subscripts for conciseness. Equivalently, λ_{max} is equal to the minimum of

$$\mathbb{E}\left[\operatorname{Var}\left(\xi(A) \,|\, Y \right) \right] + \mathbb{E}\left[s_\eta(Y)\xi(A) \right]' H_\eta^{-1} \mathbb{E}\left[s_\eta(Y)\xi(A) \right]$$

subject to $\mathbb{E}\xi(A) = 1$ and $\operatorname{Var}\xi(A) \leq 1$.

The first-order conditions of the corresponding Lagrangian are

$$2\xi(a)\pi_\gamma(a) - 2\mathbb{E}\left[\mathbb{E}\left(\xi(A) \,|\, Y \right) \,|\, a \right] \pi_\gamma(a) + 2\mathbb{E}\left[s_\eta(Y) \,|\, a \right]' H_\eta^{-1} \mathbb{E}\left[s_\eta(Y)\xi(A) \right] \pi_\gamma(a)$$
$$+ \lambda_1 \pi_\gamma(a) + 2\lambda_2 \xi(a)\pi_\gamma(a) = 0.$$

Hence, denoting $\nu(a) = \xi(a) - 1$,

$$\mathbb{E}\left[\mathbb{E}\left(\nu(A) \,|\, Y \right) \,|\, a \right] - \mathbb{E}\left[s_\eta(Y) \,|\, a \right]' H_\eta^{-1} \mathbb{E}\left[s_\eta(Y)\nu(A) \right] = (1 + \lambda_2)\nu(a).$$

Note that, since at the solution $\operatorname{Var}\nu(A) = 1$, we have $\lambda_{max} = 1 + \lambda_2$. It thus follows that λ_{max} is the maximum eigenvalue of the operator

$$\nu(a) \mapsto \mathbb{E}\left[\mathbb{E}\left(\nu(A) \,|\, Y \right) \,|\, a \right] - \mathbb{E}\left[s_\eta(Y) \,|\, a \right]' H_\eta^{-1} \mathbb{E}\left[s_\eta(Y)\nu(A) \right].$$

Note that in the known (β, γ) case this operator is equal to $\mathbb{H}_{\mathcal{A}}$, whereas in the estimated (β, γ) case it is a projected version of $\mathbb{H}_{\mathcal{A}}$.

We can thus approximate λ_{max} by the maximum eigenvalue of the following $S \times S$ matrix computed by simulation, the (s_1, s_2) element of which is (using the notation of Subsection 4.3)

$$\sum_{\tau=1}^{S} \frac{g_\beta(Y^{(\tau)} \,|\, A^{(s_1)}) g_\beta(Y^{(\tau)} \,|\, A^{(s_2)})}{\left(\sum_{s'=1}^{S} g_\beta(Y^{(\tau)} \,|\, A^{(s')})\right)^2}$$
$$- \left(\sum_{\tau=1}^{S} \frac{d_\eta(Y^{(\tau)}) g_\beta(Y^{(\tau)} \,|\, A^{(s_1)})}{\sum_{s'=1}^{S} g_\beta(Y^{(\tau)} \,|\, A^{(s')})}\right)' \left(\sum_{\tau=1}^{S} d_\eta(Y^{(\tau)}) d_\eta(Y^{(\tau)})'\right)^{-1} \left(\sum_{\tau=1}^{S} \frac{d_\eta(Y^{(\tau)}) g_\beta(Y^{(\tau)} \,|\, A^{(s_2)})}{\sum_{s'=1}^{S} g_\beta(Y^{(\tau)} \,|\, A^{(s')})}\right).$$

This matrix is equal to $G'QG = \widetilde{G}'\widetilde{G}$.

B.4 Two additional semi-parametric examples

In this subsection of the appendix we analyze two additional semi-parametric examples: a potential outcomes model under selection on observables and a demand model.

B.4.1 Average treatment effects under selection on observables

In our first example we consider a setting with a binary treatment variable D, and two potential outcomes $Y(0), Y(1)$ which we assume to be independent of D given a vector X of covariates (e.g., Rosenbaum and Rubin, 1983b). Our target parameter is the average treatment effect $\delta = \mathbb{E}(Y(1) - Y(0))$.

Let $\pi = f_d(y \,|\, x)$ denote the density of $Y(d)$ given $X = x$, for $d \in \{0, 1\}$. We assume that the propensity score $p(x) = \Pr(D = 1 \,|\, X = x)$ is correctly specified. However, we allow the reference parametric specification π_γ, where $\gamma = (\gamma_0, \gamma_1)$, to be misspecified. We focus on a regression specification for $\mathbb{E}_\gamma(Y(d) \,|\, X) = X'\gamma_d$, and assume that under the reference model $Y(d)$ is normally distributed given $X = x$ with variance σ^2. The value of σ^2 has no impact on the analysis. While $\frac{1}{n}\sum_{i=1}^{n} X_i'(\gamma_1 - \gamma_0)$ is consistent for δ under correct specification of the conditional means, it is generally inconsistent otherwise. In the analysis we treat the propensity score $p(x)$ and the parameter γ as known.

Given a function $h(y, d, x)$, we consider the estimator of δ given by $\widehat{\delta}_{h,\gamma} = \frac{1}{n}\sum_{i=1}^{n} X_i'(\gamma_1 - \gamma_0) + \frac{1}{n}\sum_{i=1}^{n} h(Y_i, D_i, X_i)$. The analysis differs slightly from the setup of Section 4, due to the presence of the *two* densities f_0 and f_1. We rely on the Kullback-Leibler divergence

$D_{KL}(f_0 f_1, \widetilde{f}_0 \widetilde{f}_1)$ between products of densities in order to define neighborhoods. Using similar arguments as in Section 4 we find

$$b_\epsilon(h,\gamma) = \frac{}{\epsilon^{\frac{1}{2}} \sqrt{\widehat{\mathrm{Var}}_\gamma \left(Y(1) - X'\gamma_1 - p(X)h(Y(1),1,X)\right) + \widehat{\mathrm{Var}}_\gamma \left(Y(0) - X'\gamma_0 + (1-p(X))h(Y(0),0,X)\right)},}$$

and

$$h_\epsilon^{\mathrm{MMSE}}(y,d,x,\gamma) = \frac{d(y - x'\gamma_1)}{p(x) + (\epsilon n)^{-1}} + \frac{(1-d)(y - x'\gamma_0)}{1 - p(x) + (\epsilon n)^{-1}}.$$

The minimum-MSE estimator of the average treatment effect is thus

$$\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \frac{1}{n}\sum_{i=1}^n X_i'(\gamma_1 - \gamma_0) + \frac{1}{n}\sum_{i=1}^n \frac{D_i(Y_i - X_i'\gamma_1)}{p(X_i) + (\epsilon n)^{-1}} + \frac{(1-D_i)(Y_i - X_i'\gamma_0)}{1 - p(X_i) + (\epsilon n)^{-1}}.$$

Notice that as ϵ tends to infinity $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ becomes

$$\lim_{\epsilon\to\infty} \widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \frac{1}{n}\sum_{i=1}^n X_i'(\gamma_1 - \gamma_0) + \frac{1}{n}\sum_{i=1}^n \frac{D_i(Y_i - X_i'\gamma_1)}{p(X_i)} + \frac{(1-D_i)(Y_i - X_i'\gamma_0)}{1 - p(X_i)},$$

which is closely related to the inverse propensity weighting estimator, and is consistent irrespective of whether the conditional means are correctly specified, provided $0 < p(X) < 1$ with probability one. The term $(\epsilon n)^{-1}$ provides a regularization which guarantees that the minimum-MSE estimator remains well-behaved in the absence of such overlap.

B.4.2 A demand model

In our second example we consider a demand setting with J products. Individual i chooses product $Y_i = j$ if j maximizes her utility $U_{ij} = X_{ij}'\beta_j + A_{ij}$, where X_{ij} are observed characteristics and A_{ij} are random preference shocks; that is,

$$Y_i = j \;\Leftrightarrow\; X_{ij}'\beta_j + A_{ij} \geq X_{ik}'\beta_k + A_{ik} \text{ for all } k \neq j. \tag{B14}$$

We assume that the vector of individual preference shocks $A = (A_1, ..., A_J)$ is independent of $X = (X_1, ..., X_J)$, with density π. We are interested in predictions from the demand model, such has counterfactual market shares under different prices or other attributes of the goods. We denote such effects as $\delta_{\theta_0} = \mathbb{E}_{\theta_0}\left(\Delta(A, X, \beta_0)\right)$, for a known function Δ, where θ_0 denotes the true value of $\theta = (\beta, \pi)$.

We start with a reference parametric specification $\theta(\eta) = (\beta, \pi_\gamma)$ for $\eta = (\beta, \gamma)$. A common example of a reference specification is A_j being i.i.d. type-I extreme value, leading

to a multinomial logit demand model. Note that in this particular case π is parameter-free. A widely echoed concern in the literature on demand analysis is that properties of the logit, in particular independence of irrelevant alternatives (IIA), may have undesirable consequences for the estimation of δ_{θ_0}; see Anderson *et al.* (1992), for example.

Assuming that β and γ are known for simplicity, in this example we have, by (36) and (38),

$$b_\epsilon(h,\beta,\gamma) = \epsilon^{\frac{1}{2}} \sqrt{\widehat{\mathbb{E}}_\gamma \left[\left(\Delta(A,X,\beta) - \mathbb{E}_\gamma \Delta(\widetilde{A},X,\beta) - \sum_{j=1}^{J} q_j(A,X,\beta) h(j,X) \right)^2 \right]},$$

where

$$q_j(a,x,\beta) = \mathbf{1}\left\{ x_j'\beta_j + a_j \geq x_k'\beta_k + a_k \text{ for all } k \neq j \right\}.$$

Moreover, we have, for all $k = 1, ..., K$ and x,

$$\mathbb{E}_{\beta,\gamma}\left[\sum_{j=1}^{J} q_j(A,x,\beta) h_\epsilon^{\text{MMSE}}(j,x,\beta) \;\middle|\; Y = k, X = x \right] + (\epsilon n)^{-1} h_\epsilon^{\text{MMSE}}(k,x,\beta)$$
$$= \mathbb{E}_{\beta,\gamma}\left[\Delta(A,x,\beta) \;\middle|\; Y = k, X = x \right] - \mathbb{E}_\gamma \Delta(A,x,\beta).$$

B.5 Individual effects in panel data (continued)

In this subsection we consider panel data models where $g_\beta(y \,|\, a, x)$ may be misspecified. Let us start with the case where neither g_β nor π_γ are correctly specified. We treat β and γ as known for simplicity. We have

$$b_\epsilon(h,\beta,\gamma) = \epsilon^{\frac{1}{2}} \sqrt{\widehat{\text{Var}}_{\beta,\gamma}\Big[\Delta(A,X) - h(Y,X) \Big]}.$$

In this case, there is a unique h function which minimizes the bias (to first-order), which corresponds to the *empirical Bayes* h function; that is,

$$h^{\text{EB}}(y,x,\beta,\gamma) = \mathbb{E}_{\beta,\gamma}\big[\Delta(A,X) \,|\, Y = y, X = x\big] - \mathbb{E}_\gamma\big[\Delta(A,X) \,|\, X = x\big], \quad \text{for all } y, x.$$

Note that here there is no scope for achieving fixed-T or even large-T identification (except in the trivial case where $\Delta(A,X) = \Delta(X)$ does not depend on A).

Consider next the case where π_γ is correctly specified, but g_β may be misspecified. We have

$$b_\epsilon(h,\beta,\gamma) = \epsilon^{\frac{1}{2}} \sqrt{\widehat{\text{Var}}_{\beta,\gamma}\Big[\Delta(A,X) - h(Y,X) - \mathbb{E}_\beta\left[\Delta(A,X) - h(\widetilde{Y},X) \,|\, A, X \right] \Big]}.$$

C Extensions

In this section of the appendix we study several extensions of our approach. We start by considering models defined by moment restrictions, and we then outline various other generalizations.

C.1 Models defined by moment restrictions

In this subsection we consider a model where the parameter θ_0 does not fully determine the distribution f_0 of Y, but satisfies the system of moment conditions (48). This system may be just-identified, over-identified or under-identified. We focus on asymptotically linear GMM estimators that satisfy (49) for an η-specific parameter vector $a(\eta)$. We assume that the remainder in (49) is uniformly bounded similarly as in (20). In this case local robustness with respect to η takes the form

$$\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{f_0} \nabla_\eta \Psi(Y, \theta(\eta))\, a(\eta) = 0. \tag{C15}$$

It is natural to focus on asymptotically linear GMM estimators here, since f_0 is unrestricted except for the moment condition (48).

To derive the worst-case bias of $\widehat{\delta}$ note that, by (48), for any $\eta \in \mathcal{B}$ and any $\theta_0 \in \Gamma_\epsilon(\eta)$ we have

$$\mathbb{E}_{f_0} \Psi(Y, \theta(\eta)) = -\left[\mathbb{E}_{f_0} \nabla_\theta \Psi(Y, \theta(\eta))\right]' (\theta_0 - \theta(\eta)) + o(\epsilon^{\frac{1}{2}}),$$

so, under appropriate regularity conditions,

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \left| \mathbb{E}_{f_0} \widehat{\delta} - \delta_{\theta_0} \right| = \epsilon^{\frac{1}{2}} \left\| \nabla_\theta \delta_{\theta(\eta)} + \mathbb{E}_{f_0} \nabla_\theta \Psi(Y, \theta(\eta))\, a(\eta) \right\|_\eta + o(\epsilon^{\frac{1}{2}}) + o(n^{-\frac{1}{2}}).$$

The worst-case MSE of

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

is thus

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

To obtain an explicit expression for the minimum-MSE estimator, let us focus on the case where θ_0 is finite-dimensional and $\|\cdot\|_\eta = \|\cdot\|_{\Omega^{-1}}$. Let us define

$$V_{\theta(\eta)} = \mathbb{E}_{f_0} \Psi(Y, \theta(\eta)) \Psi(Y, \theta(\eta))', \quad K_{\theta(\eta)} = \mathbb{E}_{f_0} \nabla_\theta \Psi(Y, \theta(\eta)), \quad K_\eta = \mathbb{E}_{f_0} \nabla_\eta \Psi(Y, \theta(\eta)).$$

For all $\eta \in \mathcal{B}$ we aim to minimize

$$\epsilon \left\|\nabla_\theta \delta_{\theta(\eta)} + K_{\theta(\eta)} a(\eta)\right\|^2_{\Omega^{-1}} + a(\eta)' \frac{V_{\theta(\eta)}}{n} a(\eta), \qquad \text{subject to } \nabla_\eta \delta_{\theta(\eta)} + K_\eta a(\eta) = 0.$$

A solution is given by[27]

$$\begin{aligned} a_\epsilon^{\text{MMSE}}(\eta) = &-B^\dagger_{\theta(\eta),\epsilon} K'_\eta \left(K_\eta B^\dagger_{\theta(\eta),\epsilon} K'_\eta\right)^{-1} \nabla_\eta \delta_{\theta(\eta)} \\ &- B^\dagger_{\theta(\eta),\epsilon} \left(I - K'_\eta \left(K_\eta B^\dagger_{\theta(\eta),\epsilon} K'_\eta\right)^{-1} K_\eta B^\dagger_{\theta(\eta),\epsilon}\right) K'_{\theta(\eta)} \Omega^{-1} \nabla_\theta \delta_{\theta(\eta)}, \end{aligned} \tag{C16}$$

where $B_{\theta(\eta),\epsilon} = K'_{\theta(\eta)} \Omega^{-1} K_{\theta(\eta)} + (\epsilon n)^{-1} V_{\theta(\eta)}$, and $B^\dagger_{\theta(\eta),\epsilon}$ is its Moore-Penrose generalized inverse. Note that, in the likelihood case and taking $\Psi(y,\theta) = \nabla_\theta \log f_\theta(y)$, the function $h(y,\eta) = a_\epsilon^{\text{MMSE}}(\eta)' \Psi(y, \theta(\eta))$ simplifies to (28).

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

$$a_0^{\text{MMSE}}(\eta) = -V^\dagger_{\theta(\eta)} K'_\eta \left(K_\eta V^\dagger_{\theta(\eta)} K'_\eta\right)^{-1} \nabla_\eta \delta_{\theta(\eta)}.$$

In this case the minimum-MSE estimator

$$\widehat{\delta}_\epsilon^{\text{MMSE}} = \delta_{\theta(\widehat{\eta})} + a_0^{\text{MMSE}}(\widehat{\eta})' \frac{1}{n} \sum_{i=1}^n \Psi(Y_i, \theta(\widehat{\eta}))$$

is the one-step approximation to the optimal GMM estimator based on the reference model, given a preliminary estimator $\widehat{\eta}$. To obtain a feasible estimator one simply replaces the expectations in $V_{\theta(\eta)}$ and K_η by sample analogs.

As a second special case, consider ϵ tending to infinity. Focusing on the known-η case for simplicity, $a_\epsilon^{\text{MMSE}}(\eta)$ tends to

$$-\underbrace{\left(V^\dagger_{\theta(\eta)}\right)^{1/2} \left[\left(V^\dagger_{\theta(\eta)}\right)^{1/2} K'_{\theta(\eta)} \Omega^{-1} K_{\theta(\eta)} \left(V^\dagger_{\theta(\eta)}\right)^{1/2}\right]^\dagger \left(V^\dagger_{\theta(\eta)}\right)^{1/2} K'_{\theta(\eta)} \Omega^{-1}}_{=K^{\text{ginv}}_{\theta(\eta)}} \nabla_\theta \delta_{\theta(\eta)},$$

where $K^{\text{ginv}}_{\theta(\eta)}$ is a generalized inverse of $K_{\theta(\eta)}$, 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

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

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

Example. Consider again the OLS/IV example of Subsection 3.2, but now drop the Gaussian assumptions on the distributions. For known Π, the set of moment conditions corresponds to the moment functions

$$\Psi(y,x,z,\theta) = \begin{pmatrix} x(y - x'\beta - \rho'(x - \Pi z)) \\ z(y - x'\beta) \end{pmatrix}.$$

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

$$K_\eta = -\mathbb{E}_{f_0}(XW'),\ K_{\theta(\eta)} = -\mathbb{E}_{f_0}\begin{pmatrix} XX' & XZ' \\ (X - \Pi Z)X' & 0 \end{pmatrix},\ V_{\theta(\eta)} = \mathbb{E}_{f_0}\left((Y - X'\beta)^2 WW'\right).$$

Given a preliminary estimator $\widetilde{\beta}$, $V_{\theta(\eta)}$ can be estimated as $\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widetilde{\beta})^2 W_i W_i'$, whereas K_η and $K_{\theta(\eta)}$ can be estimated as sample means. The estimator based on (C16) then interpolates nonlinearly between the OLS and IV estimators, similarly as in the likelihood case.

Remarks. If the researcher is willing to specify a complete parametric model f_{θ_0} compatible with the moment conditions (48), the choice of ϵ can then be based on the approach described in Subsection 2.5. Alternatively, the choice of ϵ can be based on specification testing ideas which do not require full specification, such as a test of exogeneity in the OLS/IV example above.

Lastly, the approach outlined here can be useful in fully specified structural models when the likelihood function, score and Hessian of the model are difficult to compute. Given a set of moment conditions implied by the structural model, instead of implementing (28) one may compute the optimal a vector though (C16), which only involves the moment functions and their derivatives. When the moments are computed by simulation, their derivatives can be approximated using numerical differentiation. Note that this minimum-MSE estimator has a different interpretation (and a larger mean squared error) compared to the estimator in (28) that relies on the full likelihood structure.

C.2 Bayesian interpretation

A different approach to account for misspecification of the reference model would be to specify a prior on the parameter θ_0. A Bayesian decision maker could then compute the posterior mean $\mathbb{E}[\delta_{\theta_0} \mid Y_1, \ldots, Y_n]$. As we discuss in C.2.1 below, in the parametric case of Section 3, when θ_0 is endowed with the Gaussian prior $\mathcal{N}(\theta(\eta), \epsilon\Omega^{-1})$ and η is endowed with

a non-dogmatic prior, this posterior mean coincides with our minimum-MSE estimator up to smaller-order terms; that is,

$$\mathbb{E}\left[\delta_{\theta_0} \,|\, Y_1, \ldots, Y_n\right] = \widehat{\delta}_{\epsilon}^{\text{MMSE}} + o_P(\epsilon^{\frac{1}{2}}) + o_P\left(n^{-\frac{1}{2}}\right). \tag{C17}$$

A related question is the interpretation of our minimax estimator in terms of a least-favorable prior distribution. As we discuss in C.2.2 below, in the parametric case a least-favorable prior for θ_0 given η concentrated on the neighborhood $\Gamma_\epsilon(\eta)$ puts all mass at the boundary of $\Gamma_\epsilon(\eta)$.

C.2.1 Gaussian prior

Consider the known η case to start with. To see that (C17) holds, note that, under sufficient regularity conditions,

$$\mathbb{E}\left[\delta_{\theta_0} \,|\, Y_1, \ldots, Y_n, \eta\right] = \delta_{\theta(\eta)} + (\nabla_\theta \delta_{\theta(\eta)})' \mathbb{E}\left[\theta_0 - \theta(\eta) \,|\, Y_1, \ldots, Y_n, \eta\right] + o_P(\epsilon^{\frac{1}{2}}), \tag{C18}$$

where

$$\begin{aligned}\mathbb{E}\left[\theta_0 - \theta(\eta) \,|\, Y_1, \ldots, Y_n, \eta\right] &= \frac{\int (\theta_0 - \theta(\eta)) \prod_{i=1}^n f_{\theta_0}(Y_i) \exp\left(-\frac{1}{2\epsilon}(\theta_0 - \theta(\eta))'\Omega(\theta_0 - \theta(\eta))\right) d\theta_0}{\int \prod_{i=1}^n f_{\theta_0}(Y_i) \exp\left(-\frac{1}{2\epsilon}(\theta_0 - \theta(\eta))'\Omega(\theta_0 - \theta(\eta))\right) d\theta_0} \\ &= \epsilon^{\frac{1}{2}} \frac{\int u \prod_{i=1}^n f_{\theta(\eta)+\epsilon^{\frac{1}{2}}u}(Y_i) \exp\left(-\frac{1}{2}u'\Omega u\right) du}{\int \prod_{i=1}^n f_{\theta(\eta)+\epsilon^{\frac{1}{2}}u}(Y_i) \exp\left(-\frac{1}{2}u'\Omega u\right) du}.\end{aligned}$$

Now, since, up to smaller terms,

$$\prod_{i=1}^n f_{\theta(\eta)+\epsilon^{\frac{1}{2}}u}(Y_i) \approx \prod_{i=1}^N f_{\theta(\eta)}(Y_i) \exp\left(\epsilon^{\frac{1}{2}} u' \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) - \frac{1}{2}\epsilon n u' H_{\theta(\eta)} u\right),$$

we have

$$\begin{aligned}&\mathbb{E}\left[\theta_0 - \theta(\eta) \,|\, Y_1, \ldots, Y_n, \eta\right] \\ &= \epsilon^{\frac{1}{2}} \frac{\int u \exp\left(\epsilon^{\frac{1}{2}} u \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) - \frac{1}{2}u'[\Omega + \epsilon n H_{\theta(\eta)}]u\right) du}{\int \exp\left(\epsilon^{\frac{1}{2}} u \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) - \frac{1}{2}u'[\Omega + \epsilon n H_{\theta(\eta)}]u\right) du} + o_P(\epsilon^{\frac{1}{2}}) + o_P\left(n^{-\frac{1}{2}}\right) \\ &= \epsilon n [\Omega + \epsilon n H_{\theta(\eta)}]^{-1} \frac{1}{n} \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) + o_P(\epsilon^{\frac{1}{2}}) + o_P\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

Lastly, in the case where η is estimated, let us endow it with a non-dogmatic prior. Under regularity conditions, taking expectations in (C18) with respect to the posterior distribution of η implies that (C17) holds.

C.2.2 Least favorable prior

Consider the known η case, in the parametric setting with weighted Euclidean norm. Consider the minimax problem

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

where ρ belongs to a class of priors supported on $\Gamma_\epsilon(\theta(\eta))$.

Assuming that the order of the infimum and supremum can be reversed, a least-favorable prior ρ^{LF} solves

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

For given h the integral is equal to

$$\begin{aligned}
&\int_{\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[(\widehat{\delta}_{h,\eta} - \delta_{\theta_0})^2\right] \rho(\theta_0) d\theta_0 \\
&= \int_{\Gamma_\epsilon(\eta)} \left(\frac{\mathrm{Var}_{\theta(\eta)}\, h(Y,\eta)}{n} + \left(\delta_{\theta(\eta)} + \mathbb{E}_{\theta_0} h(Y,\eta) - \delta_{\theta_0}\right)^2 \right) \rho(\theta_0) d\theta_0 + o(\epsilon) + o\left(n^{-1}\right) \\
&= \frac{\mathrm{Var}_{\theta(\eta)}\, h(Y,\eta)}{n} \\
&+ \left(\mathbb{E}_{\theta(\eta)} h(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)}\right)' \Omega^{-\frac{1}{2}} V_\Omega(\rho) \Omega^{-\frac{1}{2}} \left(\mathbb{E}_{\theta(\eta)} h(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)}\right) \\
&+ o(\epsilon) + o\left(n^{-1}\right),
\end{aligned}$$

where

$$V_\Omega(\rho) = \int_{\Gamma_\epsilon(\eta)} \Omega^{\frac{1}{2}} \left(\theta_0 - \theta(\eta)\right) \left(\theta_0 - \theta(\eta)\right)' \Omega^{\frac{1}{2}} \rho(\theta_0) d\theta_0.$$

This quantity (net of the lower-order terms) is minimized, subject to the unbiasedness restriction, at h^* which solves

$$h^*(y,\eta) = n\, \nabla_\theta \log f_{\theta(\eta)}(y)' \Omega^{-\frac{1}{2}} V_\Omega(\rho) \Omega^{-\frac{1}{2}} \left(\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h^*(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y)\right).$$

Let now

$$v = \Omega^{-1} \left(\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h_\epsilon^{\mathrm{MMSE}}(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y)\right),$$

and consider a prior ρ^{LF} that puts all mass at $\theta(\eta) + \epsilon^{\frac{1}{2}} v / \|v\|_\Omega$, say. Note that ρ^{LF} puts all mass at the boundary of $\Gamma_\epsilon(\eta)$ (see also footnote 7).

Then

$$V_\Omega(\rho^{\mathrm{LF}}) = \epsilon\, \frac{\Omega^{\frac{1}{2}} v v' \Omega^{\frac{1}{2}}}{v' \Omega v}.$$

Moreover, it can be checked that, for $\rho = \rho^{\mathrm{LF}}$,

$$h^*(\cdot, \eta) = h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta),$$

and that ρ^{LF} is least-favorable.

In the case where η is estimated, consider the following problem, for a given prior w on η and a preliminary estimator $\widehat{\eta}$,

$$\inf_h \sup_\rho \int_{\mathcal{B}} \int_{\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[(\widehat{\delta}_{h,\widehat{\eta}} - \delta_{\theta_0})^2 \right] \rho(\theta_0 \,|\, \eta) w(\eta) d\theta_0 d\eta,$$

where $\rho(\cdot \,|\, \eta)$ belongs to a class of priors supported on $\Gamma_\epsilon(\theta(\eta))$ for all η. Note that this formulation provides a Bayesian interpretation for the weight function w appearing in (19).

Applying the above arguments to the estimated-η case, one can derive a related least-favorable prior that satisfies

$$V_\Omega(\rho^{\mathrm{LF}}(\cdot|\eta)) = \epsilon \frac{\Omega^{\frac{1}{2}} v v' \Omega^{\frac{1}{2}}}{v' \Omega v}, \quad \text{for} \quad v = \Omega^{-1} \left(\widetilde{\nabla}_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h_\epsilon^{\mathrm{MMSE}}(Y, \eta) \widetilde{\nabla}_\theta \log f_{\theta(\eta)}(Y) \right).$$

For such a prior, the implied optimal $h^*(\cdot, \eta)$ is again equal to $h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta)$.

C.3 Partial identification

Here we discuss how our approach relates to a partial identification analysis. We focus on the general setup described in Section 2, for a given reference model indexed by a known η. Consider the following *restricted identified set* for δ_{θ_0}, where f_0 denotes the population distribution of Y,

$$\mathcal{S}_{\epsilon,\eta} = \{ \delta_{\theta_0} \,:\, \theta_0 \in \Theta,\, f_{\theta_0} = f_0,\, d(\theta_0, \theta(\eta)) \leq \epsilon \}.$$

$\mathcal{S}_{\epsilon,\eta}$ is equal to the intersection of the identified set for δ_{θ_0} with the image by δ of the neighborhood $\Gamma_\epsilon(\eta)$.

Proposition C3. *For any $\epsilon \geq 0$ we have*

$$\operatorname{diam} \mathcal{S}_{\epsilon,\eta} \leq 2 \inf_h b_\epsilon(h, \eta), \tag{C19}$$

where $\operatorname{diam} \mathcal{S}_{\epsilon,\eta} = \sup_{(\delta_1,\delta_2) \in \mathcal{S}_{\epsilon,\eta}^2} |\delta_2 - \delta_1|$ denotes the diameter of the restricted identified set, and the infimum is taken over any function h such that $\mathbb{E}_{f_0} h(Y)$ exists. Moreover, (C19) holds with equality whenever

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y, \eta) = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} - \left(\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y, \eta) \right) = b_\epsilon(h, \eta). \tag{C20}$$

Note that (C20) is satisfied when $\Gamma_\epsilon(\eta)$ is symmetric around $\theta(\eta)$ and $\delta_{\theta_0} - \mathbb{E}_{\theta_0} h(Y, \eta)$ is linear in θ_0. In addition, (C20) approximately holds – up to lower-order terms – when ϵ tends to zero.

Proof. Let h such that $\mathbb{E}_{f_0} h(Y)$ exists. Let $(\delta_1, \delta_2) \in \mathcal{S}^2_{\epsilon,\eta}$, with $\delta_1 = \delta_{\theta_1}$ and $\delta_2 = \delta_{\theta_2}$. Then $\mathbb{E}_{\theta_1} h(Y) = \mathbb{E}_{\theta_2} h(Y) = \mathbb{E}_{f_0} h(Y)$, so

$$\begin{aligned}|\delta_2 - \delta_1| &= |\delta_{\theta_2} - \delta_{\theta_1}| \\ &\leq |\delta_{\theta_2} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_2} h(Y) + \mathbb{E}_{\theta(\eta)} h(Y)| + |\delta_{\theta_1} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_1} h(Y) + \mathbb{E}_{\theta(\eta)} h(Y)| \leq 2 b_\epsilon(h, \eta).\end{aligned}$$

This shows (C19).

To see when (C19) holds with equality, note that the problem

$$\sup_{(\delta_1,\delta_2)\in\mathcal{S}^2_{\epsilon,\eta}} \delta_{\theta_2} - \delta_{\theta_1}$$

can equivalently be written as

$$\sup_{(\theta_1,\theta_2)\in\Gamma_\epsilon(\eta)^2} \delta_{\theta_2} - \delta_{\theta_1} + \int_{\mathcal{Y}} \lambda_1(y) f_{\theta_1}(y) dy + \int_{\mathcal{Y}} \lambda_2(y) f_{\theta_2}(y) dy, \tag{C21}$$

where λ_1 and λ_2 are the functional Lagrange multipliers associated with the restrictions $f_{\theta_1} = f_0$ and $f_{\theta_2} = f_0$, respectively. Hence, (C21) is equal to

$$\begin{aligned}&\sup_{\theta_1\in\Gamma_\epsilon(\eta)} \left(-\delta_{\theta_1} + \delta_{\theta(\eta)} + \int_{\mathcal{Y}} \lambda_1(y) f_{\theta_1}(y) dy\right) + \sup_{\theta_2\in\Gamma_\epsilon(\eta)} \left(\delta_{\theta_2} - \delta_{\theta(\eta)} + \int_{\mathcal{Y}} \lambda_2(y) f_{\theta_2}(y) dy\right) \\ &\quad = b_\epsilon(\lambda_1, \eta) + b_\epsilon(-\lambda_2, \eta) \geq 2 \inf_h b_\epsilon(h, \eta),\end{aligned}$$

where we have used (C20).

∎

C.4 Different approaches

Distance function. Consider again the setup of Section 3, now equipped with the distance measure $d(\theta_0, \theta) = (\max_{k=1,\ldots,\dim\theta} |\theta_k - \theta_{0k}|)^2$. In this case,

$$\|u\|_{\eta,\epsilon} = \|u\|_\eta = \sum_{k=1}^{\dim\theta} |u_k|$$

is the ℓ^1 norm of the vector u. Hence, computing $h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta)$ in (11) requires minimizing a convex function which combines a quadratic objective function with an ℓ^1 penalty, similarly as in the LASSO (Tibshirani, 1996).

Choice of epsilon. While in the paper we focus on a model detection error approach as in Hansen and Sargent (2008), other rules could be used to set ϵ. For example, an alternative calibration strategy is to target a maximal percentage increase in variance relative to the estimate based on the parametric reference model. Specifically, one may set $\epsilon(k)$ such that the variance of $\widehat{\delta}^{\text{MMSE}}_{\epsilon(k)}$ is lower than k times the variance of $\delta_{\theta(\widehat{\eta}^{\text{MLE}})}$, for any given constant $k \geq 1$, where $\widehat{\eta}^{\text{MLE}}$ is the MLE based on the reference model. If k is kept fixed as n tends to infinity, ϵn will be constant in the limit. For example, in the parametric case of Section 3, by (28) and given a preliminary estimator $\widehat{\eta}$, $\epsilon = \epsilon(k)$ can be chosen such that:

$$(\widetilde{\nabla}_\theta \delta_{\theta(\widehat{\eta})})'[\widetilde{H}_{\theta(\widehat{\eta})} + (\epsilon n)^{-1}\Omega]^{-1}\widetilde{H}_{\theta(\widehat{\eta})}[\widetilde{H}_{\theta(\widehat{\eta})} + (\epsilon n)^{-1}\Omega]^{-1}\widetilde{\nabla}_\theta \delta_{\theta(\widehat{\eta})} = (k-1)\,(\nabla_\eta \delta_{\theta(\widehat{\eta})})' H_\eta^{-1} \nabla_\eta \delta_{\theta(\widehat{\eta})}.$$

Role of the unbiasedness constraint (2). The asymptotic unbiasedness restriction (2) on the candidate h functions is motivated by the aim to focus on an estimator which performs well under the reference model, while in addition providing some robustness away from the reference model. Interestingly, in the case with known η and a weighted Euclidean norm, (29) remains valid when (2) is dropped. In this case our minimax objective coincides with a minimax regret criterion.

Loss function. While we focus on a quadratic loss function other losses are compatible with our approach. In fact, for any loss function $L(a, b)$ that is strictly convex and smooth in its first argument, minimizing the maximum value of

$$\mathbb{E}_{\theta_0}\left[L\left(\widehat{\delta}_{h,\widehat{\eta}}, \delta_{\theta_0}\right)\right]$$

on Γ_ϵ will lead to the same expressions for the minimum-MSE h function. This is due to our focus on a local asymptotic approach, and the fact that $L(a, b) \approx c|a - b|^2$ when $|a - b| \approx 0$.

Fixed-ϵ bias. In this paper we rely on a small-ϵ asymptotic. The tractability of our results relies crucially on a local approach. Nevertheless, in some models it is possible to provide relatively simple bias formulas for fixed ϵ. To see this, let us consider the setup of Section 4 for known β and γ. We have the following result.

Proposition C4. *For any $\epsilon > 0$ we have*

$$b_\epsilon(h, \beta, \gamma) = \left| C\mathbb{E}_\gamma\left[\left(\widetilde{\Delta}_\gamma(A, \beta) - \mathbb{E}_\beta\left(h(Y) \mid A\right)\right)\exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\gamma(A, \beta) - \mathbb{E}_\beta\left(h(Y) \mid A\right)\right)\right)\right]\right|, \tag{C22}$$

for $\widetilde{\Delta}_\gamma(a,\beta) = \Delta(a,\beta) - \mathbb{E}_\gamma\Delta(A,\beta)$*, and* $C > 0$ *and* λ_2 *two constants which satisfy equations (C23)-(C24) given in the proof.*

Proposition C4 provides an explicit expression for the bias, for *any* $\epsilon > 0$. Note that both C and λ_2 depend on ϵ. When ϵ tends to zero one can show that $1/\lambda_2$ tends to zero, and the bias converges to the expression in (36).

While it would be theoretically possible to follow a fixed-ϵ approach throughout the analysis, instead of the local approach we advocate, proceeding in that way would face several challenges. First, the bias in (C22) depends on parameters C and λ_2 which need to be recovered given ϵ, increasing computational cost. Second, simple fixed-ϵ derivations seem to be limited to settings where the parameter θ_0 (that is, π_0 in the present setting) enters the likelihood function linearly. Under linearity, similar derivations have been used in other contexts, see Schennach (2013) for an example. The third and main challenge is that characterizing mean squared errors and confidence intervals would become less tractable, while as we have seen those remain simple calculations under a local approximation. Lastly, note that the local approach allows us to provide insights into the form of the solution, as shown by our discussion of the panel data example.

Proof. Let us omit the reference to β, γ for conciseness, and denote $\pi = \pi_\gamma$. Consider the maximization of $\left|\delta_{\pi_0} - \delta_\pi - \int h(y) f_{\pi_0}(y) dy\right|$ with respect to π_0. Let $\widetilde{\Delta}_\pi(a) = \Delta(a) - \delta_\pi$. The corresponding Lagrangian is

$$\mathcal{L} = \iint_{\mathcal{Y}\times\mathcal{A}} \left(\widetilde{\Delta}_\pi(a) - h(y)\right) g(y \,|\, a)\pi_0(a) dy da + \lambda_1 \int_{\mathcal{A}} \pi_0(a) da + 2\lambda_2 \int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right) \pi_0(a) da.$$

The first-order conditions with respect to π_0 are then

$$\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y) g(y \,|\, a) dy + [\lambda_1 + 2\lambda_2] + 2\lambda_2 \log\left(\frac{\pi_0(a)}{\pi(a)}\right) = 0.$$

Hence, using that π_0 integrates to one,

$$\pi_0(a) = C \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x) g(y \,|\, a) dy\right)\right) \pi(a),$$

where

$$C^{-1} = \int_{\mathcal{A}} \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x) g(y \,|\, a) dy\right)\right) \pi(a) da. \tag{C23}$$

Since, at the least-favorable π_0, $2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right)\pi_0(a)da = \epsilon$, we have

$$\epsilon = 2\log C - \frac{C}{\lambda_2}\int_{\mathcal{A}}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x)g(y\,|\,a)dy\right)\times \\ \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x)g(y\,|\,a)dy\right)\right)\pi(a)da. \tag{C24}$$

It follows that

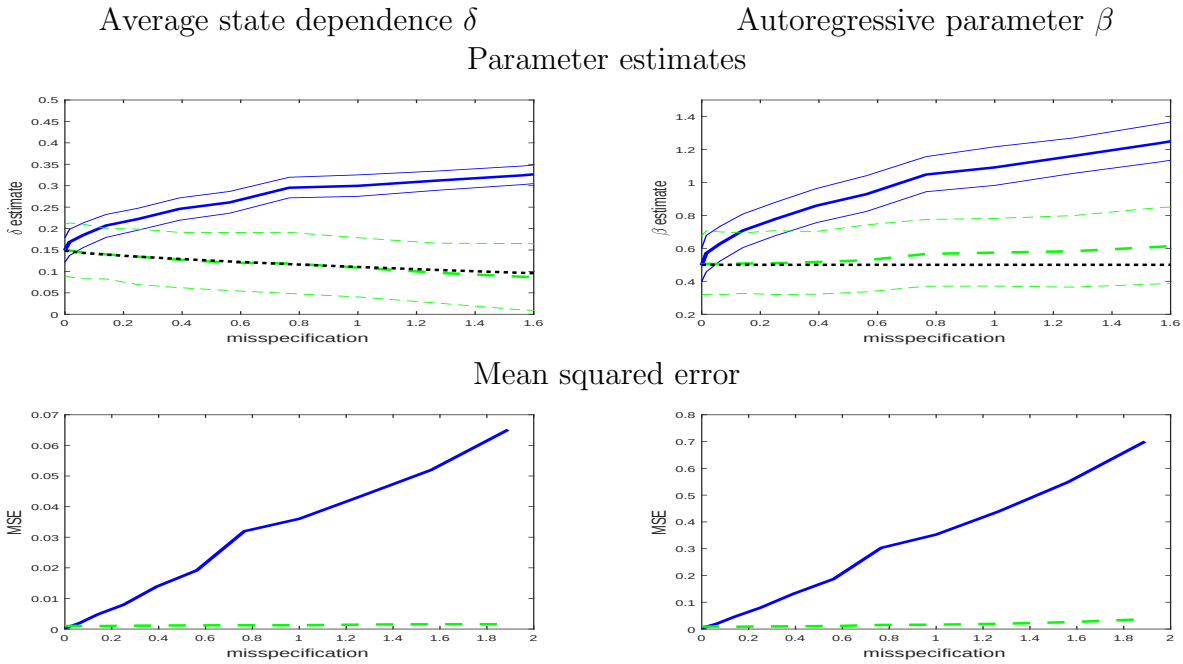
$$b_\epsilon(h) = \left|C\int_{\mathcal{A}}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x)g(y\,|\,a)dy\right)\times \right. \\ \left. \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x)g(y\,|\,a)dy\right)\right)\pi(a)da\right|,$$

where C and λ_2 satisfy (C23)-(C24).

Hence (C22) follows.

∎

Figure C1: Estimates and mean squared error of random-effects and minimum-MSE estimators under varying amount of misspecification, $p = 10^{-10}$



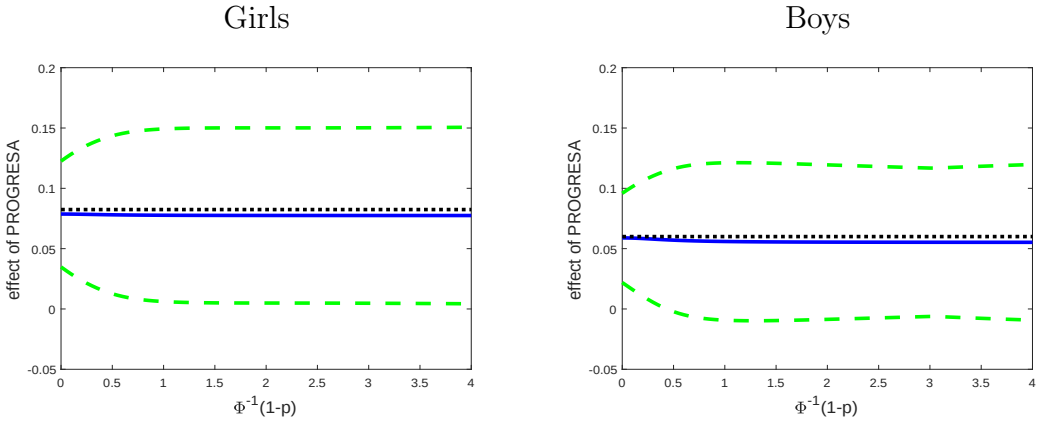
Notes: Random-effects (solid) and minimum-MSE (dashed) for δ (left graphs) and β (right graphs). True parameter values are shown in dotted. $n = 500$, $T = 5$. The reference specification for π is normal with mean $-.25 + .5Y_{i0}$ and standard deviation .8, whereas the true π_0 is normal with the same standard deviation and mean $-.25+\nu+.5Y_{i0}$. On the x-axis we report twice the KL divergence; that is, $\nu^2/.64$. Top panel: mean and 95% interval. Bottom panel: mean squared error. ϵ is chosen according to (43) for a detection error probability $p = 10^{-10}$. (μ, σ) are treated as known.

Table C1: Effect of the PROGRESA subsidy and counterfactual reforms, reference model estimated on both controls and treated

	Model-based		Minimum-MSE		Experimental	
	PROGRESA impacts					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.082	.060	.078	.055	.087	.050
non-robust CI	(.026,.139)	(.018,.102)	-	-	-	-
robust CI	(-.012,.177)	(-.058,.178)	(.005,.150)	(-.008,.119)	-	-
	Counterfactual 1: doubling subsidy					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.154	.112	.147	.105	-	-
robust CI	(-.008,.315)	(-.091,.315)	(.025,.270)	(-.004,.214)	-	-
	Counterfactual 2: unconditional transfer					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.007	.000	.003	-.012	-	-
robust CI	(-.542,.557)	(-.478,.478)	(-.201,.207)	(-.193,.169)	-	-

Notes: Sample from Todd and Wolpin (2006). $p = .01$. CI are 95% confidence intervals. The unconditional transfer amounts to 5000 pesos in a year.

Figure C2: Effect of the PROGRESA subsidy as a function of the detection error probability, reference model estimated on both controls and treated



Notes: Sample from Todd and Wolpin (2006). $\epsilon(p)$ is chosen according to (32), with $\Phi^{-1}(1-p)$ 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. Girls (left) and boys (right).