Minimizing Sensitivity to Model Misspecification*

Stéphane Bonhomme† Martin Weidner‡

August 9, 2018

Abstract

We propose a framework to improve the predictions based on an economic model, and the estimates of the model parameters, 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, which allows us to perform systematic sensitivity analysis in both point-identified and partially-identified settings. 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.

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, 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- 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 away from a *reference specification* of the model. This specification 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 away 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 the worst-case mean squared error (MSE) in a given neighborhood of the reference model. This worst case is influenced by the directions of model misspecification which matter most for the parameter of interest. We focus 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 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 use 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 Kolesar, 2016).

Our local approach leads to tractable expressions for bias and mean squared error as well as for the minimum-mean squared error estimators in a given neighborhood of the reference model. Minimum-mean squared error estimators generically take 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, between the OLS estimator and the IV estimator. When the rank condition is satisfied, 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 (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 (Bonhomme, 2012, Bonhomme and Davezies, 2017) the estimator converges to a semi-parametrically consistent estimator as the product of the neighborhood size and the sample size tends to infinity. Moreover, our approach remains informative when identification is irregular or point-identification fails, as we illustrate in a numerical exercise based on a dynamic probit model.

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, and 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), which account 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, while the robust statistics literature has mostly focused on data contamination and fully nonparametric classes.

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), 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 overidentified 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 *targeted* forms of 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) and Fessler and Kasy (2018). 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 outline the main features of our approach. In Sections 3 and 4 we detail the approach in parametric and semi-parametric 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 present several extensions in Section 7, and we conclude in Section 8.

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 Model and estimators

We observe a random sample $(Y_i : i = 1, \ldots, n)$ from the distribution $f_\theta(y) = f(y \mid \theta)$, where $\theta \in \Theta$ is a finite or infinite dimensional parameter. The parameter of interest is δ_θ, a function or functional of θ, which is assumed to be scalar. 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 any 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 example 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 sees 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.

In the local asymptotic framework we study, 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 variance calculations involve considerable mathematical difficulties, a small-ϵ analysis allows us to rely on linearization techniques and obtain simple, closed-form 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 the following sense,

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

Equation (2) is a form of local regularity of the estimator (e.g., Bickel *et al.*, 1993).

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,

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

where we write Y to denote Y_i for one representative $i \in \{1, \ldots, n\}$. Under (3), the estimator $\widehat{\delta}$ is asymptotically unbiased for the target parameter $\delta_{\theta(\eta)}$ under the reference model. Second, h is *locally robust* with respect to η in the following sense,

$$\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{\theta(\eta)} \nabla_\eta h(Y, \eta) = 0, \quad \text{for all } \eta \in \mathcal{B}, \tag{4}$$

where ∇_η is the derivative operator. The constraint (4) 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 (3) with respect to η we obtain the following equivalent expression for (4),

$$\mathbb{E}_{\theta(\eta)}\, h(Y, \eta)\, \nabla_\eta \log f_{\theta(\eta)}(Y) = \nabla_\eta \delta_{\theta(\eta)}, \quad \text{for all } \eta \in \mathcal{B}. \tag{5}$$

Local robustness (4)-(5) follows from properties of influence functions under general conditions (see, e.g., Chernozhukov *et al.*, 2016).

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) and (2), for a suitable h function satisfying (3) and (4)-(5). In Appendix A we provide more detail about the asymptotically linear representation (1), and we give a number of examples of estimators where it holds.[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 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

$$\widehat{\delta}_{h,\widehat{\eta}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \widehat{\eta}) \tag{6}$$

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 (4) that, under weak conditions on the preliminary estimator $\widehat{\eta}$, $\widehat{\delta}_{h,\widehat{\eta}}$ satisfies (2) 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 β based on the reference specification, for example, the OLS estimator under the assumption that ρ (the coefficient of an omitted control variable) is zero. Such an estimator will be consistent and efficient under the reference model. However, under misspecification 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 ("empirical Bayes") estimator is obtained by substituting an integral with respect to the

[1]Note that, in (1) and (2), the estimator is expanded around the reference value $\delta_{\theta(\eta)}$. As we discuss in Appendix A, these 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).

posterior distribution of individual effects based on $\pi_{\widehat{\gamma}}$. Both estimators are consistent, and the random-effects estimator is efficient, under the parametric reference specification in fixed-lengths panels. However they are generally biased under misspecification, that is, when 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 Bias and mean squared error when the reference model is known

For presentation purposes, in this subsection 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 $\mathcal{B} = \{\eta\}$. In the next subsection we generalize the analysis to allow η to be estimated by a preliminary estimator $\widehat{\eta}$, which is a relevant setting in many applications.

For any $\epsilon \geq 0$, and any $\eta \in \mathcal{B}$, 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

$$\|u\|_{\eta,\epsilon} = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \epsilon^{-\frac{1}{2}}\, u(\theta_0 - \theta(\eta)), \quad \|u\|_\eta = \lim_{\epsilon \to 0} \|u\|_{\eta,\epsilon}\,. \tag{7}$$

We assume that the distance measure d is chosen such that $\|\cdot\|_\eta$ is unique and well-defined, and 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 (3) 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

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \left| \mathbb{E}_{\theta_0} \widehat{\delta} - \delta_{\theta_0} \right| = b_\epsilon(h, \eta) + o(\epsilon^{\frac{1}{2}}) + o(n^{-\frac{1}{2}}), \tag{8}$$

where

$$b_\epsilon(h, \eta) = \epsilon^{\frac{1}{2}} \left\| \nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)}\, h(Y, \eta)\, \nabla_\theta \log f_{\theta(\eta)}(Y) \right\|_\eta, \tag{9}$$

for $\|\cdot\|_\epsilon$ the dual norm defined in (7), where ∇_θ is a Gâteaux derivative when θ is infinite-dimensional. Then, the worst-case MSE in $\Gamma_\epsilon(\eta)$ can be expanded as follows, again under

appropriate regularity conditions,

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta} - \delta_{\theta_0}\right)^2\right] = b_\epsilon(h,\eta)^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n} + o(\epsilon) + o\left(n^{-1}\right). \tag{10}$$

In order to construct estimators with minimum worst-case MSE we define, for any function h satisfying (3),

$$\widehat{\delta}_{h,\eta} = \delta_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i,\eta). \tag{11}$$

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

$$h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta) = \underset{h(\cdot,\eta)}{\operatorname{argmin}}\ \epsilon\ \left\|\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)}\, h(Y,\eta)\ \nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n}$$
$$\text{subject to (3).} \tag{12}$$

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 (10) only needs to be calculated under the reference model, and the optimization problem (12) is convex.

Note that, for $\epsilon = 0$ we have $\delta_0^{\mathrm{MMSE}} = \delta_{\theta(\eta)}$, 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]

Examples (cont.). 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

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

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 common 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

[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.

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,

$$\|q\|_\eta = \sqrt{\mathrm{Var}_{\theta(\eta)}\left(q(A)\right)}. \tag{13}$$

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

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

It thus follows that

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

Hence (9) and (12) become, respectively,

$$b_\epsilon(h,\eta) = \epsilon^{\frac{1}{2}} \sqrt{\mathrm{Var}_{\theta(\eta)}\left(\Delta(A) - \mathbb{E}\left[h(Y,\eta) \,|\, A\right]\right)}, \tag{14}$$

and

$$h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta) = \underset{h(\cdot,\eta)}{\mathrm{argmin}}\ \epsilon\, \mathrm{Var}_{\theta(\eta)}\left(\Delta(A) - \mathbb{E}\left[h(Y,\eta) \,|\, A\right]\right) + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n}$$
$$\text{subject to (3).} \tag{15}$$

As in the parametric case, the MSE-minimization problem (15) is thus quadratic in h, and computing $h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta)$ amounts to solving a quadratic problem.

2.3 Bias and mean squared error when the reference model is estimated

We now consider the case where the parameter η is unknown. As in the previous subsection we focus on estimators $\widehat{\delta}$ which satisfy expansion (1). In addition to (3), we assume that the function h satisfies the local robustness condition (5). For given $\eta \in \mathcal{B}$, the worst-case bias and MSE of $\widehat{\delta}$ in $\Gamma_\epsilon(\eta)$ can then be expanded for small ϵ, similarly as in (8) and (10). This allows one to compare different estimators in terms of their worst-case bias and MSE.

To construct minimum-MSE estimators, 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 (6). We search for functions $h(\cdot,\eta)$ satisfying both (3) and (5) which minimize an asymptotic approximation to the following integrated worst-case MSE,

$$\int_{\mathcal{B}} \left\{ \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_{h,\widehat{\eta}} - \delta_{\theta_0} \right)^2 \right] \right\} w(\eta) d\eta, \tag{16}$$

where w is a positive 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, since the asymptotic expansion (10) of the MSE and the restrictions (3) and (5) only involve $h(\cdot,\eta)$ at a given η value, and since $\widehat{\delta}_{h,\widehat{\eta}}$ satisfies (2) under local robustness, the minimization of (16) decouples asymptotically and its solution does not depend on the weight function w.

As a result, minimizing a small-ϵ approximation to (16) with respect to $\{h(\cdot,\eta) : \eta \in \mathcal{B}\}$ simply amounts to solving the following program, for all $\eta \in \mathcal{B}$,

$$h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta) = \underset{h(\cdot,\eta)}{\operatorname{argmin}} \; \epsilon \left\| \nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h(Y,\eta) \, \nabla_\theta \log f_{\theta(\eta)}(Y) \right\|_\eta^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n}$$
$$\text{subject to (3) and (5)}, \tag{17}$$

where we note that (17) is again a convex optimization problem. We then define the minimum-MSE estimator of δ_{θ_0} as

$$\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n} \sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i, \widehat{\eta}). \tag{18}$$

In practice, (17) only needs to be solved at $\eta = \widehat{\eta}$. In addition, since $\widehat{\delta}_{h,\widehat{\eta}}$ satisfies (2) the form of the minimum-MSE estimator is not affected by the choice of the preliminary estimator $\widehat{\eta}$.

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

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

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

$$h_0^{\mathrm{MMSE}}(y,\eta) = \left[\nabla_\eta \log f_{\theta(\eta)}(y) \right]' H_\eta^{-1} \, \nabla_\eta \delta_{\theta(\eta)}. \tag{19}$$

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]

Another interesting special case of the minimum-MSE h function arises in the limit $\epsilon \to \infty$,[4] 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

$$\lim_{\epsilon \to \infty} h_{\epsilon}^{\text{MMSE}}(y, \eta) = \left[\nabla_{\theta} \log f_{\theta(\eta)}(y)\right]' H_{\theta(\eta)}^{-1} \nabla_{\theta} \delta_{\theta(\eta)}. \tag{20}$$

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. Thus, if θ_0 is identified, then the estimator $\widehat{\delta}_{\epsilon}^{\text{MMSE}}$, for any ϵ, is an 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 (17) is only a *local* approximation to the full MSE-minimization problem.

Note that neither (19) nor (20) involve the particular choice of distance function with respect to which neighborhoods are defined. The minimum-MSE estimator interpolates, nonlinearly in general, between these two estimators. For given $\epsilon > 0$ the minimum-MSE estimator will depend on the chosen distance function.

The estimator in (20) is "orthogonalized" or "locally robust" (e.g., Neyman, 1959, Chernozhukov *et al.*, 2016) with respect to the large parameter θ.[5] 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. In 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.

[3] Such one-step approximations are classical estimators in statistics; see for example Bickel *et al.* (2013, pp. 43–45).

[4] Equivalently, the same limiting quantity is attained if ϵ is kept fixed as $n \to \infty$, or if ϵn tends to infinity.

[5] 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.

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 (2), (3) and (5). Let $\mu \in (0,1)$ be a confidence level. Using (2) and the expression for $b_\epsilon(h,\eta)$ in (9) we have, as n tends to infinity and ϵn tends to a constant,

$$\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \Pr_{\theta_0}\left[\left|\widehat{\delta}-\delta_{\theta_0}\right| \leq b_\epsilon(h,\eta) + \frac{\sigma_h(\theta_0,\eta)}{\sqrt{n}} c_{1-\mu/2}\right] \geq 1-\mu+o(1), \tag{21}$$

where $b_\epsilon(h,\eta)$ is given by (9), $\sigma_h^2(\theta_0,\eta) = \mathrm{Var}_{\theta_0}(h(Y,\eta))$, and $c_{1-\mu/2} = \Phi^{-1}(1-\mu/2)$ is the $(1-\mu/2)$-standard normal quantile.

Let us define the following interval

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

where $\widehat{\sigma}_h^2$ is the sample variance of $h(Y_1,\widehat{\eta}),\ldots,h(Y_n,\widehat{\eta})$. Under suitable smoothness conditions on $b_\epsilon(h,\eta)$ and $\sigma_h(\theta_0,\eta)$, it follows that 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; see Appendix C for a formal statement. 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 Kolesar, 2016).[6]

2.5 Choice of ϵ

Confidence intervals, like 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. Specifically, for a fixed probability $p \in (0,1)$ and given a parameter η, let $\epsilon = \epsilon(p,\eta)$ be such that

$$\inf_{\theta_0\in\Gamma_\epsilon(\eta)} \Pr_{\theta(\eta)}\left(\sum_{i=1}^n \log\left(\frac{f_{\theta_0}(Y_i)}{f_{\theta(\widehat{\eta})}(Y_i)}\right) > 0\right) = p + o(1). \tag{23}$$

Taking ϵ according to (23) guarantees that, for some θ_0 in (the d-closure of) $\Gamma_\epsilon(\eta)$, the probability of incorrectly detecting that θ_0 is more likely to have generated the data than $\theta(\eta)$ is approximately equal to p. Achieving a lower p requires setting a higher ϵ.

[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 $|\mathcal{N}(1, \frac{\widehat{\sigma}_h^2}{b_\epsilon(h,\widehat{\eta})^2 n})|$.

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

$$\sup_{\theta_0 \in \Gamma_\epsilon(\widehat{\eta})} (\theta_0 - \theta(\widehat{\eta}))' \widetilde{H}_{\theta(\widehat{\eta})} (\theta_0 - \theta(\widehat{\eta})) = \frac{4 \left(\Phi^{-1}(p)\right)^2}{n}, \tag{24}$$

where $\widetilde{H}_{\theta(\eta)} = H_{\theta(\eta)} - H_{\theta(\eta)} G_\eta' H_\eta^{-1} G_\eta H_{\theta(\eta)}$, for $G_\eta = \nabla_\eta \theta(\eta)'$ (a $\dim \theta \times \dim \eta$ matrix). We will see that the implied ϵ value has a closed-form expression as a function of p in the parametric and semi-parametric models we will analyze in the next two sections.

Setting $\epsilon = \epsilon(p)$ according to (24) 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 "harder to detect" based on a sample of n observations, for example through specification testing. 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 using the Hansen-Sargent strategy, as we do, provides an interpretable metric to assess how "large" or "small" deviations are.

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 of any given estimator to possible misspecification of the reference model. In addition, our approach delivers a collection of minimum-MSE estimators $\widehat{\delta}^{\,\mathrm{MMSE}}_{\epsilon(p)}$ for different p. In practice, it can be informative to report the full curve $\widehat{\delta}^{\,\mathrm{MMSE}}_{\epsilon(p)}$ as a function of p, along with the estimator corresponding to a preferred p level. We will report such quantities in our empirical illustration in Section 6.

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 Section 7 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 we study the case where θ is finite-dimensional and the distance function 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 function is based on $\|\cdot\|_\Omega$, the small-ϵ approximation to the bias of $\widehat{\delta}$ is given by (9), 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 (17) is quadratic and its solution satisfies

$$\begin{aligned} h_\epsilon^{\mathrm{MMSE}}(y,\eta) = &\left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} \\ &+ (\epsilon n)\left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \Omega^{-1} \left(\widetilde{\nabla}_\theta \delta_{\theta(\eta)} - \mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(Y)\right]\right), \end{aligned} \tag{25}$$

where $\widetilde{\nabla}_\theta = \nabla_\theta - H_{\theta(\eta)} G_\eta' H_\eta^{-1} \nabla_\eta$ is a projected gradient operator, and where 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

$$\begin{aligned} h_\epsilon^{\mathrm{MMSE}}(y,\eta) = &\left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} \\ &+ \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \left[\widetilde{H}_{\theta(\eta)} + (\epsilon n)^{-1}\Omega\right]^{-1} \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \end{aligned} \tag{26}$$

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 $h_0^{\mathrm{MMSE}}(y,\eta)$ given by (19), the minimum-MSE function $h_\epsilon^{\mathrm{MMSE}}(\cdot,\eta)$ provides a further adjustment that is motivated by robustness concerns.

Here we have derived the expression of the minimum-MSE estimator by minimizing an asymptotic approximation to the worst-case MSE. In Appendix D we compare the integrated worst-case mean squared error (16) of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ to that of any estimator $\widehat{\delta}$ satisfying (2) for a suitable h function for which (3) and (4)-(5) hold. We show that, under mild regularity conditions on the model and the weight function $w(\eta)$, the minimum-MSE estimator has smaller integrated worst-case mean squared error, up to terms of smaller asymptotic order. See Corollary D1 in Appendix D.

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 (20),

$$\lim_{\epsilon\to\infty} h_\epsilon^{\text{MMSE}}(y,\eta) = \left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1}\, \nabla_\eta \delta_{\theta(\eta)} + \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \; \widetilde{H}^\dagger_{\theta(\eta)} \; \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \quad (27)$$

where $\widetilde{H}^\dagger_{\theta(\eta)}$ denotes the Moore-Penrose generalized inverse of $\widetilde{H}_{\theta(\eta)}$.[7] Comparing (27) and (26) 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 (24), which here simplifies to

$$\sup_{v\in\mathbb{R}^{\dim\theta}\,:\,v'\Omega v\le\epsilon} v'\widetilde{H}_{\theta(\widehat{\eta})}v = \frac{4\left(\Phi^{-1}(p)\right)^2}{n}, \quad (28)$$

the solution of which is

$$\epsilon(p) = \frac{4\left(\Phi^{-1}(p)\right)^2}{n\cdot\lambda_{max}(\Omega^{-\frac{1}{2}}\widetilde{H}_{\theta(\widehat{\eta})}\Omega^{-\frac{1}{2}})}, \quad (29)$$

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 (28), 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).

Remark 1: parameters of the reference model. Robust estimation of specific parameters of the reference model, such as structural parameters in economic models, can be analyzed within our framework. Consider the first example we outlined in Section 2, where $\theta = (\beta, \rho)$, $\eta = \beta$, $\theta(\eta) = (\beta, 0)$, and $\delta_\theta = c'\beta$ for some known vector c. In addition, take Ω to be block-diagonal with β-block Ω_β and ρ-block Ω_ρ. By (26), and making use of the fact

[7] In fact, $\widetilde{H}^\dagger_{\theta(\eta)}$ in (27) can be replaced by any generalized inverse of $\widetilde{H}_{\theta(\eta)}$.

that the minimum-MSE h function satisfies (5), we have

$$h_\epsilon^{\text{MMSE}}(y,\beta) = \left[\nabla_\beta \log f_{\beta,0}(y)\right]' H_\beta^{-1} c \\ - \left[\nabla_\rho \log f_{\beta,0}(y) - H_{\rho\beta}H_\beta^{-1}\nabla_\beta \log f_{\beta,0}(y)\right]' \left[H_\rho - H_{\rho\beta}H_\beta^{-1}H_{\rho\beta}' + (\epsilon n)^{-1}\Omega_\rho\right]^{-1} H_{\rho\beta}H_\beta^{-1} c, \tag{30}$$

where $H_{\rho\beta} = \mathbb{E}_{\beta,0}\left[\nabla_\rho \log f_{\beta,0}(Y)\right]\left[\nabla_\beta \log f_{\beta,0}(Y)\right]'$, $H_\rho = \mathbb{E}_{\beta,0}\left[\nabla_\rho \log f_{\beta,0}(Y)\right]\left[\nabla_\rho \log f_{\beta,0}(Y)\right]'$.

Remark 2: incorporating covariates. It is common in applications with covariates to model the conditional distributions of outcomes Y given covariates X parametrically as $f_\theta(y \mid x)$, 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 minimum-MSE h function takes a similar form to the expressions derived above, except that it involves averages over the covariates sample $X_1,\ldots,X_n$. For example, (26) becomes

$$h_\epsilon^{\text{MMSE}}(y,x,\eta) = \left[\nabla_\eta \log f_{\theta(\eta)}(y \mid x)\right]' \left(\widehat{\mathbb{E}}_X H_\eta\right)^{-1} \nabla_\eta \delta_{\theta(\eta)} \\ + \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y \mid x)\right]' \left[\widehat{\mathbb{E}}_X \widetilde{H}_{\theta(\eta)} + (\epsilon n)^{-1}\Omega\right]^{-1} \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \tag{31}$$

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

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

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. For simplicity we assume that Π and σ^2 are known. The parameters are thus $\theta = (\beta, \rho)$. As a parametric 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.

Let Σ_V and Σ_X be the covariance matrices of V and X, respectively, both of which are assumed non-singular. From (30) we have[8]

$$\begin{aligned} h_\epsilon^{\text{MMSE}}(y, x, z, \beta) &= (y - x'\beta) x' \Sigma_X^{-1} \, c \\ &\quad - (y - x'\beta) \left[(x - \Pi z) - \Sigma_V \Sigma_X^{-1} x \right]' \left[\Sigma_V - \Sigma_V \Sigma_X^{-1} \Sigma_V + (\epsilon n)^{-1} \Omega_\rho \right]^{-1} \Sigma_V \Sigma_X^{-1} c. \end{aligned} \tag{32}$$

From (32) we see that, 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.[9] Indeed, it follows from (32) that

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

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

[9] 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.

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) our adjustment remains well-behaved as ϵn tends to infinity, otherwise setting a finite ϵ value is essential is order to control the increase in variance. The term $(\epsilon n)^{-1}$ in (32) 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 (29); that is,

$$\epsilon(p) = \frac{4\sigma^2 \left(\Phi^{-1}(p)\right)^2}{n \cdot \lambda_{max}\left(\Omega_\rho^{-\frac{1}{2}}\left(\Sigma_V - \Sigma_V\Sigma_X^{-1}\Sigma_V\right)\Omega_\rho^{-\frac{1}{2}}\right)}. \tag{33}$$

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 away 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 is the need to specify a model that is more flexible than the reference model in some dimensions. In parametric settings this may consist in including additional covariates, or in allowing for a more 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 and the size of that neighborhood. Our choice of ϵ is guided by the Hansen and Sargent calibration 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 apply our approach. Since we consider 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 Section 7.

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 $\mathcal{Y} \subset \mathbb{R}^m$ depends on a latent variable A supported on $\mathcal{A} \subset \mathbb{R}^q$. We denote the conditional distribution as $g_\beta(y \mid a)$, for β a finite-dimensional parameter. In turn, the distribution of A is denoted as π. The researcher postulates a parametric reference specification for π, which we denote $\pi_\gamma(a)$ 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 E we describe two additional examples: a demand model where the distributional assumptions on unobserved preference shocks may be invalid, and a treatment effects model under selection on observables where the conditional mean of potential outcomes may be misspecified.

In this setup, $\theta = (\beta, \pi)$, $\eta = (\beta, \gamma)$, and $\theta(\eta) = (\beta, \pi_\gamma)$. As a distance function on θ we use a combination of a weighted Euclidean norm on β and (twice) the Kullback-Lleibler divergence on π; that is, $d(\theta_0, \theta) = \|\beta_0 - \beta\|^2_{\Omega_\beta} + 2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right)\pi_0(a)da$. As it turns out, neither the choice of Ω_β nor the weighting of the parametric and nonparametric part in the distance function play any role in the analysis that follows.[10,11]

We first derive the form of the bias of any estimator $\widehat{\delta}$ which is asymptotically linear as in (1). It is instructive to start with the case where both β and γ are assumed to be known.

[10]In fact we obtain the same expressions in case the neighborhoods are defined in terms of the Kullback-Leibler divergence between *joint* distributions of (Y, A),

$$\widetilde{d}(\theta_0, \theta) = 2\iint_{\mathcal{Y}\times\mathcal{A}} \log\left(\frac{g_{\beta_0}(y \mid a)\pi_0(a)}{g_\beta(y \mid a)\pi(a)}\right) g_{\beta_0}(y \mid a)\pi_0(a)dyda,$$

provided $\mathbb{E}_{\beta,\gamma}\left[(\nabla_\beta \log g_\beta(Y \mid A))(\nabla_\beta \log g_\beta(Y \mid A))'\right]$ is non-singular.

[11]Moreover, except for the bound on the distance measure d, the density π_0 is left unrestricted. In particular, we do not assume that π_0 belongs a smoothness class.

In this case the small-ϵ approximation to the worst-case bias of $\widehat{\delta}$ is[12]

$$b_\epsilon(h,\beta,\gamma) = \epsilon^{\frac{1}{2}} \sqrt{\mathrm{Var}_\gamma\left(\Delta(A,\beta) - \mathbb{E}_\beta\left[h(Y)\,|\,A\right]\right)}, \tag{34}$$

where we have used that the dual of the Kullback-Leibler divergence satisfies (13), and that $\nabla_\pi \log \int_{\mathcal{A}} g_\beta(y\,|\,a)\pi(a)da = g_\beta(y\,|\,\cdot)/\int_{\mathcal{A}} g_\beta(y\,|\,a)\pi(a)da$; see also equation (14). This bias expression can be used to form confidence intervals for δ_{θ_0}, as explained in Subsection 2.4.

We now derive the form of the minimum-MSE h function, in the case of known β and γ. By (15), $h_\epsilon^{\mathrm{MMSE}}$ minimizes the following small-ϵ approximation to the MSE,

$$\epsilon\,\mathrm{Var}_\gamma\left(\Delta(A,\beta) - \mathbb{E}_\beta\left[h(Y,\beta,\gamma)\,|\,A\right]\right) + \frac{\mathrm{Var}_{\beta,\gamma}(h(Y,\beta,\gamma))}{n}, \tag{35}$$

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

$$\begin{aligned}\mathbb{E}_{\beta,\gamma}\left[\mathbb{E}_\beta(h_\epsilon^{\mathrm{MMSE}}(Y,\beta,\gamma)\,|\,A)\;|\;y\right] + (\epsilon n)^{-1}h_\epsilon^{\mathrm{MMSE}}(y,\beta,\gamma)& \\ = \mathbb{E}_{\beta,\gamma}\left[\Delta(A,\beta)\;|\;y\right] - \mathbb{E}_\gamma\Delta(A,\beta), \quad \text{for all } y\in\mathcal{Y},&\end{aligned} \tag{36}$$

where the expectations in the terms in brackets are with respect to the *posterior* distribution of the latent variable A given the outcome Y; that is,

$$p_{\beta,\gamma}(a\,|\,y) = \frac{g_\beta(y\,|\,a)\pi_\gamma(a)}{\int_{\mathcal{A}} g_\beta(y\,|\,\widetilde{a})\pi_\gamma(\widetilde{a})d\widetilde{a}}. \tag{37}$$

Note that (36) is a linear system in $h_\epsilon^{\mathrm{MMSE}}$. When Y has infinite support this system is infinite-dimensional. Since it is a Fredholm type II integral system, it can be solved uniquely given ϵn; see Carrasco *et al.* (2007), for example. In Subsection 4.3 we describe how we compute the unique minimum-MSE h function in practice.

To provide intuition about the form of the minimum-MSE h function, it is useful to write the MSE-minimization problem as a functional problem on Hilbert spaces. Indeed, minimizing the MSE is equivalent to minimizing

$$\|\Delta - \delta - \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}\,h\|_{\mathcal{A}}^2 + (\epsilon n)^{-1}\|h\|_{\mathcal{Y}}^2, \tag{38}$$

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

[12]Here and in the following, β and (β,γ) subscripts indicate that expectations and variances are taken with respect to the joint distribution (or some conditional distribution based on it) of the reference model at (β,γ).

By standard results in functional analysis (e.g., Engl *et al.*, 2000), (38) is minimized at the *regularized inverse* of the operator $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$ evaluated at $\Delta - \delta$; that is,

$$h_\epsilon^{\text{MMSE}} = \left[\mathbb{H}_\mathcal{Y} + (\epsilon n)^{-1}\mathbb{I}_\mathcal{Y}\right]^{-1}\left(\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}\,\Delta - \delta\right), \tag{39}$$

where $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$ is the conditional expectation operator given Y,[13] $\mathbb{I}_\mathcal{Y}$ is the identity operator, and $\mathbb{H}_\mathcal{Y}$ is the composition of $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$ and $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$, 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 (39) is the unique solution to (36). It is well-defined even when $\mathbb{H}_\mathcal{Y}$ is singular. The term $(\epsilon n)^{-1}$ can be interpreted as a *Tikhonov* penalization.

It is interesting to compare the minimum-MSE estimator to the posterior mean of $\Delta(A,\beta)$, which is computed according to the posterior distribution (37) with prior π_γ. By (36), the posterior mean, whose h function is $h(y) = \mathbb{E}_{\beta,\gamma}\left[\Delta(A,\beta) \mid y\right] - \mathbb{E}_\gamma\Delta(A,\beta)$, differs from the minimum-MSE estimator. Similarly as the posterior mean, our estimator updates the prior π_γ based on the data. However, the form of the update rule differs from Bayesian updating, in order to provide MSE-optimality in the neighborhood of π_γ. We further discuss the link between our approach and Bayesian approaches in Section 7.

Consider next the case where (β,γ) are estimated. Writing the first-order conditions of (17), and making use of (5), we obtain the following formula for the minimum-MSE h function,

$$\begin{aligned}&\mathbb{Q}_{\beta,\gamma}\mathbb{E}_{\beta,\gamma}\left[\mathbb{E}_\beta(h_\epsilon^{\text{MMSE}}(Y,\beta,\gamma)\,|\,A) \mid y\right] + (\epsilon n)^{-1}h_\epsilon^{\text{MMSE}}(y,\beta,\gamma)\\ &\quad= (\epsilon n)^{-1}\left[\nabla_{\beta,\gamma}\log f_{\beta,\pi_\gamma}(y)\right]' H_{\beta,\gamma}^{-1}\,\nabla_{\beta,\gamma}\,\mathbb{E}_\gamma\Delta(A,\beta) + \mathbb{Q}_{\beta,\gamma}\Big(\mathbb{E}_{\beta,\gamma}\left[\Delta(A,\beta) \mid y\right] - \mathbb{E}_\gamma\Delta(A,\beta)\Big),\end{aligned} \tag{40}$$

where $\mathbb{Q}_{\beta,\gamma}$ is the operator which projects functions of y onto the orthogonal of the score of the reference model;[14] 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 (40) is again linear in h_ϵ^{MMSE}. Note that, as in the parametric case, this expression applies in particular to the case where $\Delta(A,\beta) = \beta_k$ is a component of β.

[13] $\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}}$ is the adjoint operator to $\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}$.

[14] The orthogonality is with respect to the Hilbert space topology, with norm $\|\cdot\|_\mathcal{Y}$.

Finally, to set ϵ we rely on (29). In the case where β, γ are known this formula takes the following simple expression

$$\epsilon = \frac{4\left(\Phi^{-1}(p)\right)^2}{N}, \tag{41}$$

where we have used that here the maximum singular value of the operator $\mathbb{H}_{\mathcal{Y}}$ is equal to one.[15] Given a detection error probability p we select $\epsilon = \epsilon(p)$ according to (41). When β, γ are estimated, the maximal singular value of $\mathbb{H}_{\mathcal{Y}}$ can be approximated numerically using the approach outlined 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. We allow for $T = 1$, in which case we effectively only have one cross-section, and more generally for any fixed $T \geq 1$. 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 by $g_\beta(y \,|\, \alpha, x)$. In turn, the density of latent individual effects is denoted as $\pi(\alpha \,|\, x)$. 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

$$\delta_{\theta_0} = \mathbb{E}_{\theta_0}\left[\Delta(A, X, \beta_0)\right] = \iint_{\mathcal{A}\times\mathcal{X}} \Delta(a, x, \beta_0)\pi_0(a \,|\, x)f_X(x)dadx, \tag{42}$$

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). It is worth emphasizing that common parameters β_0 can be obtained from (42) by taking $\Delta(A, X, \beta_0) = \beta_{0k}$, for any

[15] $\mathbb{H}_{\mathcal{Y}}$ is a *doubly stochastic* operator; that is, the infinite-dimensional generalization of a doubly stochastic matrix. Hence its maximal singular value is equal to one.

component of β_0. Hence 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). Given such a parametric reference specification, an empirical counterpart to δ_{θ_0} is the *random-effects* estimator

$$\widehat{\delta}^{\mathrm{RE}} = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{A}} \Delta\left(a, X_i, \widehat{\beta}\right) \pi_{\widehat{\gamma}}(a \,|\, X_i)\, da, \tag{43}$$

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

Another commonly used estimator is the *empirical Bayes* estimator

$$\widehat{\delta}^{\mathrm{EB}} = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{A}} \Delta(a, X_i, \widehat{\beta}) \underbrace{\frac{g_{\widehat{\beta}}(Y_i \,|\, a, X_i)\pi_{\widehat{\gamma}}(a \,|\, X_i)}{\int_{\mathcal{A}} g_{\widehat{\beta}}(Y_i \,|\, \widetilde{a}, X_i)\pi_{\widehat{\gamma}}(\widetilde{a} \,|\, X_i) d\widetilde{a}}}_{= p_{\widehat{\beta},\widehat{\gamma}}(a \,|\, Y_i, X_i)} da, \tag{44}$$

where $p_{\widehat{\beta},\widehat{\gamma}}(a \,|\, Y_i, X_i)$ denotes the posterior distribution of individual effects A_i given (Y_i, X_i) implied by $g_{\widehat{\beta}}$ and $\pi_{\widehat{\gamma}}$. Both $\widehat{\delta}^{\mathrm{RE}}$ 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. Misspecifying the distribution of unobserved heterogeneity and its dependence on initial conditions and other covariates can generate severe biases (Heckman, 1981). There is an extensive literature on large-n, T properties of fixed-effects estimators of average effects (Hahn and Newey, 2004, Arellano and Hahn, 2007). In the case of random-effects estimators, Arellano and Bonhomme (2009) point out that, unlike $\widehat{\delta}^{\mathrm{RE}}$, $\widehat{\delta}^{\mathrm{EB}}$ generally remains consistent as both n and T tend to infinity when π_γ is misspecified.[16]

In this setting, our approach allows us to assess the sensitivity of various estimators of the average effect given by (42). We focus on minimizing the worst-case conditional MSE, given covariates $X_1, \ldots, X_n$. To simplify exposition, in the remainder of this section we treat the

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

parameters β and γ as known. The small-ϵ approximation to the bias of the random-effects estimator $\widehat{\delta}^{\mathrm{RE}}$ is

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

where we use "hats" to denote variances and expectations conditional on the sample of covariates.[17]

In contrast, the small-ϵ approximation to the bias of the empirical Bayes estimator $\widehat{\delta}^{\mathrm{EB}}$ 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.

Hence

$$b_\epsilon(h^{\mathrm{EB}}, \beta, \gamma) \leq b_\epsilon(h^{\mathrm{RE}}, \beta, \gamma).$$

In addition, as T tends to infinity we expect $b_\epsilon(h^{\mathrm{EB}}, \beta, \gamma)$ to tend to zero.[18] In contrast, $b_\epsilon(h^{\mathrm{RE}}, \beta, \gamma)$ is constant, independent of T. This shows that, from a fixed-T robustness perspective, the empirical Bayes estimator dominates the random-effects estimator in terms of bias. The relative ranking becomes very clear as T increases since, unlike the empirical Bayes estimator, the random-effects estimator never updates the functional form π_γ in light of the data. This finding agrees with the large-T consistency of empirical Bayes estimators and large-T *in*consistency of random-effects estimators of average effects documented in Arellano and Bonhomme (2009).

To get insight on the form of the bias and minimum-MSE h function, consider the case where there exists a function ζ such that

$$\mathbb{E}_{\beta,\gamma} \left[\zeta(Y, X) \,|\, A = a, X = x\right] = \Delta(a, x, \beta), \qquad \text{for all } a, x,$$

where we focus on the case where β is known for simplicity. The existence of such a function ζ, under suitable regularity conditions, is necessary and sufficient for δ_{θ_0} to be identified and root-n consistently estimable for fixed T (Bonhomme and Davezies, 2017). In this case it

[17] That is, $\widehat{\mathrm{Var}}_\gamma \left(\Delta(A, X, \beta)\right) = \frac{1}{n} \sum_{i=1}^n \mathrm{Var}_{\gamma \,|\, X_i} \left(\Delta(A, X_i, \beta)\right)$, with the variance inside the sum taken with respect to the conditional density $\pi_\gamma(\cdot \,|\, X_i)$.

[18] 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).

follows from (34) that $b_\epsilon(\zeta, \beta, \gamma) = 0$. Moreover, it can be shown that

$$\mathbb{E}_\gamma \Delta(A, X, \beta) + \lim_{\epsilon \to \infty} \mathbb{E}_{\theta_0} \left[h_\epsilon^{\text{MMSE}}(Y, X, \beta, \gamma) \right] = \mathbb{E}_{\theta_0} \left[\zeta(Y, X) \right] = \delta_{\theta_0}.$$

As a result, in the large-ϵ limit (and provided it has finite first moment), $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is a fixed-T, root-n consistent estimator of δ_{θ_0} *irrespective* of π_γ being correctly specified or not. In this case, in the large-ϵ limit the minimum-MSE h function thus leads to full robustness against misspecification of π_γ.

Existence of a function ζ is a strong condition, however. In particular, it presumes that δ_{θ_0} is fixed-T identified, which is not always the case. For example, parameters and average effects in discrete choice panel data models are often only partially identified (Honoré and Tamer, 2006, Chernozhukov *et al.*, 2013, Pakes and Porter, 2013). Even in the point-identified case, δ_{θ_0} may not be root-n estimable due to ill-posedness (Bonhomme and Davezies, 2017). Irrespective of whether point-identification holds or not, the minimum-MSE estimator of δ_{θ_0} is a weighted average of the random-effects estimator $\widehat{\delta}^{\text{RE}}$ and a Thikonov-regularized nonparametric estimator of δ_{θ_0}. The presence of the $(\epsilon n)^{-1}$ term in (36) can be interpreted as a *regularization*, based on the Tikhonov approach (Carrasco *et al.*, 2007). Our calibration of ϵ thus leads to a particular choice of regularization scheme. Given our choice of ϵ, which ensures that ϵn is constant in large samples, the regularized problem is well-posed. By focusing on a shrinking neighborhood of the distribution π_γ, as opposed to entertaining any possible distribution, we thus avoid ill-posedness while guaranteeing MSE-optimality within that neighborhood.[19]

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. Nevertheless, since (40) is linear in h, computing the minimum-MSE function amounts to solving a quadratic optimization problem. We proceed by simulation, drawing S values from the joint reference distribution $g_\beta \times \pi_\gamma$ of (Y, A), conditional on $X = X_i$. Given a large number of draws, an

[19] Although we abstract from common parameters in this discussion, our approach also applies to estimation of β, for example. In that case the large-ϵ limit of our minimum-MSE h function coincides, when it exists, with the influence function of a "functional differencing" estimator (Bonhomme, 2012). Note that, in contrast with the functional differencing approach, in our local approach there is no need to optimize a new objective function in order to ensure (local) robustness.

approximation to h_ϵ^{MMSE} can be computed by a simple least squares calculation. We use this approach in the numerical illustration on panel data in the next section.

5 Numerical illustration in a panel data model

5.1 Setup

In this section we present numerical simulations for the following dynamic panel data probit model with fixed-effects

$$Y_{it} = \mathbf{1}\left\{\beta Y_{i,t-1} + A_i + U_{it} \geq 0\right\}, \quad t = 2, ..., T,$$

where $U_{i2}, ..., U_{iT}$ are i.i.d. standard normal, independent of Y_{i1} and A_i. We assume that the probit conditional likelihood given individual effects and lagged outcomes is correctly specified. We specify the density $\pi = \pi_{\mu_1,\mu_2,\sigma}$ of A_i given Y_{i1}, as a Gaussian with mean $\mu_1 + \mu_2 Y_{i1}$ and variance σ^2. In this illustration we treat $\beta, \mu_1, \mu_2, \sigma$ as known parameters.

We focus on the average *state dependence* effect

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

and we consider three different estimators. The first one is the *random-effects* estimator

$$\widehat{\delta}^{\text{RE}} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{A}}\left[\Phi(\beta + a) - \Phi(a)\right]\pi(a \,|\, Y_{i1})da.$$

The second one is the *empirical Bayes* (or posterior mean) estimator

$$\widehat{\delta}^{\text{EB}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi}\left[\Phi(\beta + A_i) - \Phi(A_i) \,|\, Y_{i1}, ..., Y_{iT}\right].$$

The last one is the *minimum-MSE* estimator, for a given $\epsilon > 0$,

$$\widehat{\delta}_\epsilon^{\text{MMSE}} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{A}}\left[\Phi(\beta + a) - \Phi(a)\right]\pi(a \,|\, Y_{i1})da + \frac{1}{n}\sum_{i=1}^{n} h_\epsilon^{\text{MMSE}}(Y_{i1}, ..., Y_{iT}, \pi),$$

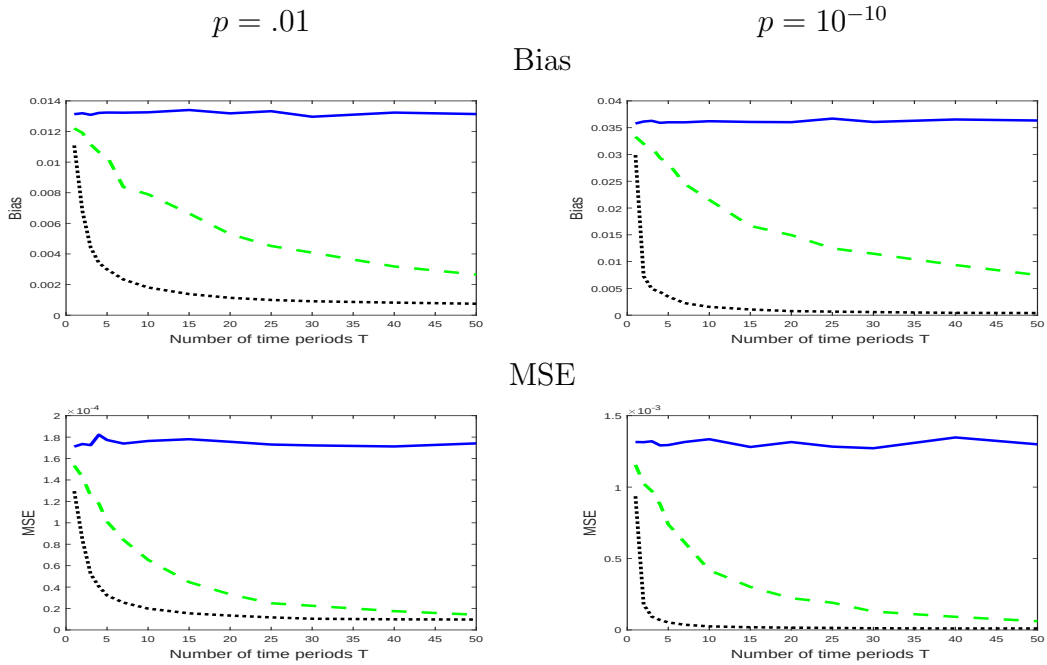
where h_ϵ^{MMSE} solves (36); that is, for all $y = (y_1, ..., y_T)$,

$$\begin{aligned}
&\mathbb{E}_\pi\left[\mathbb{E}\left(h_\epsilon^{\text{MMSE}}(y_1, Y_2, ..., Y_T, \pi) \,|\, A, y_1\right) \mid y\right] + (\epsilon n)^{-1} h_\epsilon^{\text{MMSE}}(y, \pi) \\
&\qquad = \mathbb{E}_\pi\left[\Phi(\beta + A) - \Phi(A) \mid y\right] - \int_{\mathcal{A}}\left[\Phi(\beta + a) - \Phi(a)\right]\pi(a \,|\, y_1)da,
\end{aligned}$$

where $\mathbb{E}_\pi$ is taken with respect to the posterior density of individual effects with prior π.

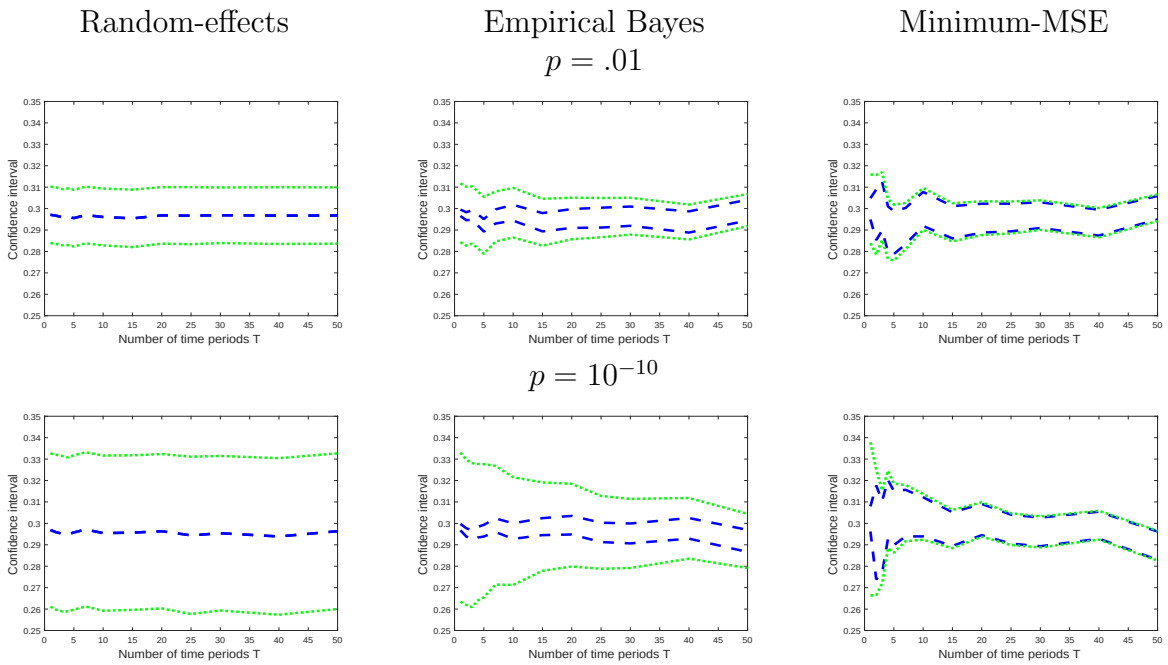
We set $\beta = 1$, $\mu_1 = -.2$, $\mu_2 = 0$, and $\sigma = .8$. In the simulation Y_{i1} is fixed to 0. We compute h_ϵ^{MMSE} by sampling from $g_\beta \times \pi$, with $S = 10,000$ support points, and solving a linear system of dimension S. We vary T between $T = 1$ and $T = 50$. The estimators are computed on a sample of size $n = 1000$.

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 $\widehat{\delta}^{\mathrm{RE}}$, the dashed line to the empirical Bayes estimator $\widehat{\delta}^{\mathrm{EB}}$, and the dotted line to the minimum-MSE estimator $\widehat{\delta}^{\mathrm{MMSE}}_{\epsilon}$. ϵ is chosen according to (41) for a detection error probability $p = .01$ (left) and $p = 10^{-10}$ (right) when $n = 1000$.

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. $n = 1000$. ϵ *is chosen according to (41) for a detection error probability* $p = .01$ *(top) and* $p = 10^{-10}$ *(bottom).*

5.2 Bias, MSE and confidence intervals

In Figure 1 we show the asymptotic bias b_ϵ and MSE for each of the three estimators, where ϵ is set according to (41) for a detection error probability $p = .01$ (left graph) and $p = 10^{-10}$ (right). On the top panel we see that the bias of the random-effects estimator (solid line) is the largest, and that it does not decrease as T grows. In contrast, the bias of the empirical Bayes 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 random-effects and empirical Bayes estimators. On the bottom panel we observe a similar relative ranking between estimators in terms of MSE.

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

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

T	1	5	10	20	50
			Bias		
Random-effects	-.0436	-.0451	-.0439	-.0447	-.0448
Empirical Bayes	-.0420	-.0342	-.0227	-.0132	-.0046
Linear probability	-	-.3101	-.1365	-.0510	-.0037
Minimum-MSE ($p = .01$)	-.0391	-.0055	-.0005	-.0033	-.0030
Minimum-MSE ($p = .10^{-10}$)	-.0388	.0032	.0020	-.0012	.0015
			Mean squared error ($\times 1000$)		
Random-effects	1.897	2.035	1.926	2.002	2.008
Empirical Bayes	1.764	1.174	.523	.182	.031
Linear probability	-	96.660	18.862	2.722	.056
Minimum-MSE ($p = .01$)	1.538	.068	.023	.027	.022
Minimum-MSE ($p = .10^{-10}$)	1.524	.081	.034	.020	.017

Notes: $n = 500$*, results for* 500 *simulations.*

for local misspecification as in (22) (dotted lines). The number of individuals is $n = 1000$, and ϵ is chosen based on (41) for a probability $p = .01$. We see that accounting for model misspecification leads to enlarged confidence intervals. However the size of the enlargement varies to a great 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 random-effects and empirical Bayes estimators widen substantially, those based on the minimum-MSE estimator remain quite informative.

5.3 Monte Carlo simulation

We next perform a Monte Carlo simulation based on the same data generating process, except that we set the population distribution of A_i to be log-normal with mean $-.2$ and standard deviation .68. The assumed distribution for A_i in the parametric reference model is still Gaussian, with the same mean and standard deviation .8. Twice the Kullback-Leibler divergence between the true log-normal density and the assumed normal density is equal to 1.52. Here our goal is to document the performance of the minimum-MSE estimator under this particular form of global misspecification.

In Table 1 we report the results of 500 Monte Carlo replications, for T ranging between 1 and 50, and $n = 500$. The upper panel shows the bias, and the lower panel shows the MSE. We report the results for five estimators: the random-effects estimator, the empirical Bayes estimator, the linear probability estimator, and the minimum-MSE estimators with ϵ set according to $p = .01$ and $p = 10^{-10}$, respectively. We see that the random-effects estimator is substantially biased and has large MSE. The linear probability estimator is severely biased in short panels in this dynamic setting. In comparison the empirical Bayes estimator has smaller bias and MSE. Moreover, the bias decreases as T increases. The minimum-MSE estimator performs best in terms of both bias and MSE, and it performs quite similarly irrespective of the choice of ϵ. Note that the calibrated neighborhood size ϵ is .04 for $p = .01$, and .32 for $p = 10^{-10}$. Hence, in both cases, the true distribution of A_i lies quite far outside of the chosen neighborhood. In spite of this, the minimum-MSE estimator performs well in this environment since it provides substantial robustness compared to random-effects, linear probability, and empirical Bayes estimators.

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 PROGRESA, 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 direct stigma effect on utility. 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. Note the direct presence of the subsidy τ in the utility function, which may reflect a stigma effect. This direct effect plays an important 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 only, under the assumption that $\lambda = 0$. In this case the parameters a, b, d can be recovered from data on control villages only. The effect of the subsidy on school attendance can then be estimated since

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

Note that data under the subsidy regime ($\tau^{treat} = \tau$) is not needed to construct empirical counterparts to such quantities. The expectations are computed in the control group, taking

advantage that 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.[20] Instead of estimating a larger model, here we adjust the predictions from the reference model against the possibility of misspecification, using data from both controls and treated. While in the present simple static context one could easily estimate a version of the model allowing for $\lambda \neq 0$, in dynamic structural models such as the one estimated by TW estimating a different model in order to assess the impact of any given form of misspecification may be computationally prohibitive. This highlights an advantage of our approach, which does not require the researcher to estimate the parameters under a new model.

To cast this setting into our framework, let $\theta = (a, b, d, \lambda)$, $\eta = (a, b, d)$, $\theta(\eta) = (a, b, d, 0)$, and: $\delta_\theta = \mathbb{E}\left(\Phi\left[a(Y+W) - (a+d)(Y+\tau) - \lambda\tau - b\right] - \Phi\left[a(Y+W) - (a+d)Y - b\right]\right)$. 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.

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 specification that in particular exploits province-level variation and variation in distance to the nearest city, similarly as in 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

[20] 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.

Table 2: Effect of the PROGRESA subsidy and counterfactual reforms

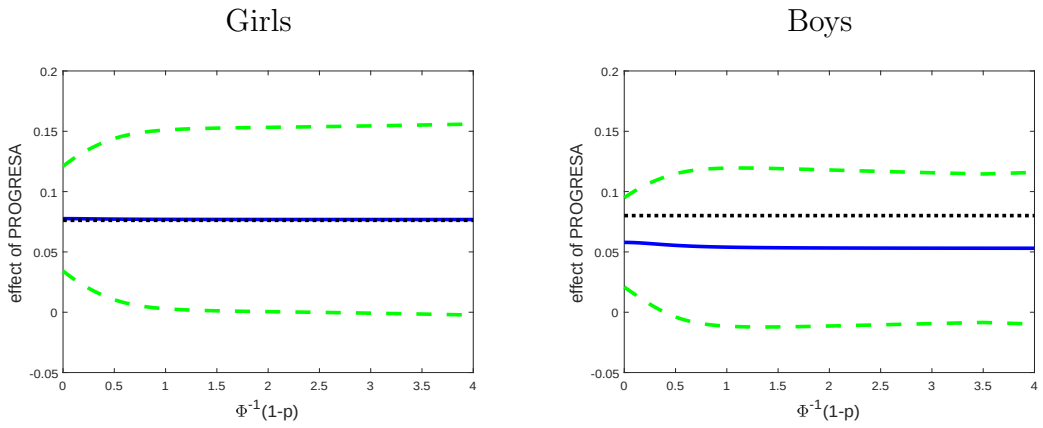
	Model-based		Minimum-MSE		Experimental	
	PROGRESA impacts					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.076	.080	.077	.053	.087	.050
non-robust CI	(.006,.147)	(.032,.129)	-	-	-	-
robust CI	(-.053,.205)	(-.062,.222)	(-.012,.166)	(-.023,.129)	-	-
	Counterfactual 1: doubling subsidy					
	Girls	Boys	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.

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 parametric 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 only. 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, 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

Figure 3: 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 (29), 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).

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 quite large.[21] Interestingly, as shown by the rightmost two columns the minimum-MSE estimates are quite close to the experimental differences in means between treated and control villages, for both genders.

When using our approach it is informative to report minimum-MSE estimates and confidence intervals for different values of the neighborhood size ϵ. In Figure 3 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 (29). 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 controls only. This suggests that, for boys, the functional form of the schooling decision is *not* invariant to treat-

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

ment 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.[22] In contrast, we find virtually 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 Appendix H.

7 Extensions

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

7.1 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 following system of moment conditions

$$\mathbb{E}_{f_0}\Psi(Y,\theta_0)=0. \tag{45}$$

This system may be just-identified, overidentified or underidentified.

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

$$\widehat{\delta}=\delta_{\theta(\eta)}+a(\eta)'\frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i,\theta(\eta))+o_P(\epsilon^{\frac{1}{2}})+o_P(n^{-\frac{1}{2}}), \tag{46}$$

for an η-specific parameter vector $a(\eta)$. We assume that the remainder in (46) is uniformly bounded similarly as in (2). In this case local robustness with respect to η is

$$\nabla_\eta\delta_{\theta(\eta)}+\mathbb{E}_{f_0}\nabla_\eta\Psi(Y,\theta(\eta))\,a(\eta)=0. \tag{47}$$

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

It is natural to focus on asymptotically linear GMM estimators here, since f_0 is unrestricted except for the moment condition (45).

To derive the worst-case bias of $\widehat{\delta}$ note that, by (45), 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

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

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

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

A solution is given by[23]

$$\begin{aligned} a_\epsilon^{\mathrm{MMSE}}(\eta) = &-B_{\theta(\eta),\epsilon}^\dagger K_\eta'\left(K_\eta B_{\theta(\eta),\epsilon}^\dagger K_\eta'\right)^{-1}\nabla_\eta\delta_{\theta(\eta)} \\ &- B_{\theta(\eta),\epsilon}^\dagger\left(I - K_\eta'\left(K_\eta B_{\theta(\eta),\epsilon}^\dagger K_\eta'\right)^{-1}K_\eta B_{\theta(\eta),\epsilon}^\dagger\right)K_{\theta(\eta)}'\Omega^{-1}\nabla_\theta\delta_{\theta(\eta)}, \end{aligned} \tag{48}$$

where $B_{\theta(\eta),\epsilon} = K_{\theta(\eta)}'\Omega^{-1}K_{\theta(\eta)} + (\epsilon n)^{-1}V_{\theta(\eta)}$, and $B_{\theta(\eta),\epsilon}^\dagger$ 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 $h(y,\eta) = a_\epsilon^{\mathrm{MMSE}}(\eta)'\Psi(y,\theta(\eta))$ simplifies to (26).

[23] Although the solution $a_\epsilon^{\mathrm{MMSE}}(\eta)$ may not be unique, the function $a_\epsilon^{\mathrm{MMSE}}(\eta)'\Psi(y,\theta(\eta))$ is unique. Here we assume that $K_\eta V_{\theta(\eta)}^\dagger K_\eta'$ is non-singular, requiring that η be identified from the moment conditions. In cases where $B_{\theta(\eta),\epsilon}$ is singular, (48) comes from the fact that $V_{\theta(\eta)}a = 0$ implies that $K_{\theta(\eta)}a = 0$ (this follows from the generalized information identity).

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

$$a_0^{\text{MMSE}}(\eta) = -V_{\theta(\eta)}^{\dagger} K_\eta' \left(K_\eta V_{\theta(\eta)}^{\dagger} K_\eta' \right)^{-1} \nabla_\eta \delta_{\theta(\eta)}.$$

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, when ϵ tends to infinity $B_{\theta(\eta),\epsilon}$ tends to $K_{\theta(\eta)}' \Omega^{-1} K_{\theta(\eta)}$, so when η is known we have

$$\lim_{\epsilon \to \infty} a_\epsilon^{\text{MMSE}}(\eta) = - \left[K_{\theta(\eta)}' \Omega^{-1} K_{\theta(\eta)} \right]^{-1} K_{\theta(\eta)}' \Omega^{-1} \nabla_\theta \delta_{\theta(\eta)},$$

in which case the minimum-MSE estimator is the one-step approximation to a GMM estimator based on the "large" model, with weight matrix Ω^{-1}. While this large-ϵ limit only exists when $K_{\theta(\eta)}$ has full column rank, for any finite $\epsilon \geq 0$ the minimum-MSE estimator remains well-defined in case of rank deficiency.[24]

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 (48) then interpolates nonlinearly between the OLS and IV estimators, similarly as in the likelihood case.

[24] Given a parameter vector a, confidence intervals can be constructed as explained in Subsection 2.4, taking $b_\epsilon(a, \widehat{\eta}) = \epsilon^{\frac{1}{2}} \left\| \nabla_\theta \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^n \nabla_\theta \Psi(Y_i, \theta(\widehat{\eta}))\, a(\widehat{\eta}) \right\|_{\Omega^{-1}}$.

Remarks. If the researcher is willing to specify a complete parametric model f_{θ_0} compatible with the moment conditions (45), 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, an alternative to implementing (26) is to compute the optimal a vector though (48), 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 (26) that relies on the full likelihood structure.

7.2 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 $h_\epsilon^{\text{MMSE}}(\cdot, \eta)$ in (12) requires minimizing a convex function which combinines a quadratic objective function with an ℓ^1 penalty, similarly as in the LASSO (Tibshirani, 1996).

Choice of epsilon. While we focus on Hansen and Sargent's (2008) model detection error approach, 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.[25]

[25] In the parametric case of Section 3, by (26) 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 (3). The asymptotic unbiasedness restriction (3) on the candidate h functions is motivated by the desire 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, (26) remains valid when (3) 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 $|a - b| \approx 0$.

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 $\mathbb{E}\left[\delta_{\theta_0} \,|\, Y_1, \ldots, Y_n\right]$. As we discuss in Appendix F, in the parametric case studied in Section 3 this posterior mean coincides with our minimum-MSE estimator up to smaller-order terms, that is,

$$\mathbb{E}\left[\delta_{\theta_0} \,|\, Y_1, \ldots, Y_n\right] = \widehat{\delta}_\epsilon^{\text{MMSE}} + o_P(\epsilon^{\frac{1}{2}}) + o_P\left(n^{-\frac{1}{2}}\right), \tag{49}$$

when θ_0 is endowed with the Gaussian prior $\mathcal{N}(\theta(\eta), \epsilon\Omega^{-1})$, and η is endowed with any non-dogmatic prior independent of θ_0.[26]

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. For fixed ϵ we have

$$b_\epsilon(h, \beta, \gamma) = \left|C\mathbb{E}_{\beta,\gamma}\left[\left(\widetilde{\Delta}_\gamma(A, \beta) - \mathbb{E}_\beta\left(h(Y) \,|\, A\right)\right)\exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\gamma(A, \beta) - \mathbb{E}_\beta\left(h(Y) \,|\, A\right)\right)\right)\right]\right|, \tag{50}$$

[26] A related question is the interpretation of our minimax estimator in terms of a least-favorable prior distribution. As we discuss in Appendix F, in the parametric case the least-favorable prior concentrated on the neighborhood $\Gamma_\epsilon(\eta)$ puts all mass at the boundary of $\Gamma_\epsilon(\eta)$.

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 (G16)-(G17) in the appendix. (50) 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 (34).

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 (50) 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 example.

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 $\Gamma_\epsilon(\eta)$.

We show in the appendix that

$$\operatorname{diam} \mathcal{S}_{\epsilon,\eta} \leq 2 \min_h b_\epsilon(h,\eta), \tag{51}$$

where $\operatorname{diam} \mathcal{S}_{\epsilon,\eta} = \sup_{(\delta_1,\delta_2)\in\mathcal{S}^2_{\epsilon,\eta}} |\delta_2 - \delta_1|$ denotes the diameter of the restricted identified set, and the minimum is taken over any function h such that $\mathbb{E}_{f_0} h(Y)$ exists. Note that (51) holds for any ϵ. Moreover, (51) holds with equality whenever

$$\sup_{\theta_0\in\Gamma_\epsilon(\eta)} \delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta) = \sup_{\theta_0\in\Gamma_\epsilon(\eta)} -\left(\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)\right) = b_\epsilon(h,\eta). \tag{52}$$

Note that (52) 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, (52) approximately holds (that is, up to lower-order terms) when ϵ tends to zero.

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 that are less sensitive to model assumptions.

Our approach can handle parametric and semi-parametric forms of misspecification. It is based on a minimax mean squared error rule, which consists of a one-step adjustment of the initial estimate. This adjustment is motivated by both robustness and efficiency, and it remains valid when the identification of the "large" model is irregular or point-identification fails. Hence, our approach provides a complement to partial identification methods, when the researcher sees her reference model as a plausible, albeit imperfect, approximation to reality.

Lastly, given a parametric reference model, implementing our estimators and confidence intervals does not require estimating a larger model. This is an attractive feature in complex models such as dynamic structural models, for which sensitivity analysis methods are needed.

References

[1] 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," unpublished manuscript.

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

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

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

[24] 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.

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

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

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

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

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

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

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

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

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

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

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

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

[37] 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.

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

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

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

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

[42] 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.

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

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

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

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

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

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

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

[50] Masten, M. A., and A. Poirier (2017): "Inference on Breakdown Frontiers," unpublished manuscript.

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

[52] Mukhin, Y. (2018): "Sensitivity of Regular Estimators." arXiv:1805.08883v1 [econ.EM].

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

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

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

[56] 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.

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

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

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

[60] 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.

[61] 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.

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

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

[64] 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.

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

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

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

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

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

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

APPENDIX

A Asymptotically linear estimators

In this section of the appendix we provide some background on the asymptotically linear representation (1), and we give several examples.

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

$$\widehat{\delta} = \delta^*_{\theta_0} + \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \theta_0) + o_{P_{\theta_0}}(n^{-\frac{1}{2}}), \tag{A1}$$

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

$$\mathbb{E}_{\theta_0}\phi(Y, \theta_0) = 0, \quad \nabla_\theta \delta^*_{\theta_0} + \mathbb{E}_{\theta_0}\nabla_\theta \phi(Y, \theta_0) = 0, \quad \text{for all } \theta_0 \in \Theta. \tag{A2}$$

The first condition in (A2) 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 inequality.[27] Expansion (A1) and conditions (A2) are satisfied for a large class of estimators, see below for examples.

We further assume that

$$\delta^*_{\theta(\eta)} = \delta_{\theta(\eta)}, \quad \text{for all } \eta \in \mathcal{B}. \tag{A3}$$

Condition (A3) 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,

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

[27] The generalized information inequality 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).

where we have used (A2) at $\theta_0 = \theta(\eta)$, and that, by (A3), $\nabla_\eta \delta^*_{\theta(\eta)} = \nabla_\eta \delta_{\theta(\eta)}$.

To relate (2), which is taken around $\delta_{\theta(\eta)}$, to expansion (A1), which is taken around $\delta^*_{\theta_0}$, note that if

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\left[\widehat{\delta} - \delta^*_{\theta_0} - \frac{1}{n}\sum_{i=1}^n \phi(Y_i,\theta_0)\right]^2 = o(n^{-1}) \tag{A4}$$

and

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\left[\delta^*_{\theta_0} + \frac{1}{n}\sum_{i=1}^n \phi(Y_i,\theta_0) - \delta^*_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^n \phi(Y_i,\theta(\eta))\right]^2 = o(\epsilon) + o(n^{-1}) \tag{A5}$$

both hold, and if (A3) is satisfied, then (2) follows immediately since we obtain

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\left[\widehat{\delta} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^n \phi(Y_i,\theta(\eta))\right]^2 = o(\epsilon) + o(n^{-1}). \tag{A6}$$

(A4) is a locally uniform extension of (A1) in an ϵ-neighborhood of the reference model, and (A5) is a stochastic equicontinuity assumption. See, e.g., Bickel *et al.* (1993) and Rieder (1994) on local asymptotic expansions of regular estimators.

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 (A4) 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 immediate to check that (A2) holds. Moreover, under suitable smoothness conditions (A5) also holds, as soon as $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\left[\frac{1}{n}\sum_{i=1}^n \nabla_\theta \phi(Y_i,\theta(\eta)) - \mathbb{E}_{\theta(\eta)}\nabla_\theta\phi(Y_i,\theta(\eta))\right]^2 = O(n^{-1})$. Lastly, (A3) 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 (A4) and (A5) hold under regularity conditions on m and q, with

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

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 $\theta_0 \in \Theta$. One can readily verify that (A2) holds. Moreover, (A3) 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 $\phi(y, \theta_0) = \delta_{\theta(\eta)} - \delta^*_{\theta_0}$. Note that (A2) and (A3) hold. Moreover, (A4) and (A5) are trivially satisfied in this case.

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 (A4) and (A5) 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 $\mathbb{E}_{\theta_0} q(Y, \eta^*_{\theta_0}) = 0$. Moreover, (A3) holds provided $\mathbb{E}_{\theta(\eta)} q(Y, \eta) = 0$ for all $\eta \in \mathcal{B}$.

B Dual of the Kullback-Leibler divergence

Let A be a random variable with domain $\mathcal{A}$, reference distribution $f_*(\alpha)$ and "true" distribution $f_0(\alpha)$. We use notation $f_*(\alpha)$ and $f_0(\alpha)$ as if those where 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 $\int_{\mathcal{A}} f(\alpha) d\alpha = 1$. 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_*.

Lemma B1 *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 $\delta_+ > 0$.[28] For $\epsilon \in (0, \delta_+^2)$ we then have*

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

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

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

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

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

where the constants $c, t \in (0, \infty)$ are determined by the constraints $\int f_0(\alpha) d\alpha = 1$ and $d(f_0, f_*) = \epsilon$. Using the constraint $\int f_0(\alpha) d\alpha = 1$ we can solve for c to obtain

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

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

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

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 $\bar{t} \in [0, t]$. ■

C Confidence intervals

In this section of the appendix we provide conditions under which the confidence interval $CI_\epsilon(1-\mu, \widehat{\delta})$ given by (22) has correct asymptotic coverage for the true parameter value δ_{θ_0}, uniformly over the neighborhood Γ_ϵ.

Assumption C1

(i) Equation (2) holds.

(ii) $\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \sigma_h(\theta_0,\eta) > 0$ *and, for all constants* a*,*

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

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

(iv) There exist a constant $C > 0$ *such that* $\sup_{\eta\in\mathcal{B}} \|\nabla_\eta b_\epsilon(h,\eta)\| \le C\epsilon^{\frac{1}{2}}$.

Theorem C1 *Let Assumption C1 hold. Then, as* n *tends to infinity and* ϵn *tends to a constant,*

$$\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \Pr_{\theta_0} \left[\delta_{\theta_0} \in CI_\epsilon(1-\mu, \widehat{\delta}) \right] \ge 1 - \mu + o(1). \tag{C7}$$

Proof. Let $(\theta_0, \eta) \in \Gamma_\epsilon$. Using (2), we have

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

where

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

for

$$\widehat{R} = \widehat{\delta} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta).$$

Now, by Assumption C1 and the Markov inequality it is easy to see that $\widehat{A} = o_p(1)$, uniformly on Γ_ϵ.

Using part (ii) in Assumption C1 then implies (C7). ■

D Optimality of the minimum-MSE estimator

In this section of the Appendix we provide conditions under which the worst-case mean squared error of the minimum-MSE estimator is smaller than that of any alternative estimator, to leading asymptotic order. Here we focus on the parametric case.

D.1 Assumptions and results

Assumption D2

(i) $f_\theta(y)$ *is four times continuously differentiable in* θ*, with derivatives uniformly bounded over* y *and* θ*.*

(ii) $\theta(\eta)$ *is twice continuously differentiable in* η *with uniformly bounded derivatives. Furthermore, we have* $\sup_{\eta\in\mathcal{B}} \left\|H_\eta^{-1}\right\| = O(1)$*.*

(iii) δ_θ *is twice continuously differentiable in* θ *with uniformly bounded derivatives.*

(iv) The estimator $\widehat{\eta}$ *satisfies* $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} \left\|\widehat{\eta} - \eta\right\|^2 = O(n^{-1})$*.*

(v) We consider $\|u\|_{\eta,\epsilon} = \|u\|_\eta = \|u\|_{\Omega^{-1}}$ *in the following, where* Ω *is a positive definite* $\dim\theta \times \dim\theta$ *matrix.*

Let $\widetilde{H}^\dagger_{\theta(\eta)}$ be the Moore-Penrose inverse of $\widetilde{H}_{\theta(\eta)}$, and define

$$\begin{aligned} Q_\epsilon^{\text{opt}}(\eta) &:= \min_{v\in\mathbb{R}^{\dim\theta}} \left[\epsilon \left\|v - \widetilde{\nabla}_\theta \delta_{\theta(\eta)}\right\|_\eta^2 + \frac{1}{n}\, v'\, \widetilde{H}^\dagger_{\theta(\eta)}\, v\right] + n^{-1} \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1}\, \nabla_\eta \delta_{\theta(\eta)} \\ &= \epsilon \left(\widetilde{\nabla}_\theta \delta_{\theta(\eta)}\right)' \left(\Omega + n\epsilon\, \widetilde{H}_{\theta(\eta)}\right)^{-1} \left(\widetilde{\nabla}_\theta \delta_{\theta(\eta)}\right) + n^{-1} \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1}\, \nabla_\eta \delta_{\theta(\eta)}. \end{aligned}$$

Theorem D2 *Let Assumption D2 hold. We then have, for* $n \to \infty$ *and* $\epsilon \to 0$*,*

$$\sup_{\eta\in\mathcal{B}} \left| \sup_{\theta_0\in\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\theta_0}\right)^2\right] - Q_\epsilon^{\text{opt}}(\eta) \right| = O\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon^{\frac{1}{2}}\right).$$

Remark: Theorem D2 provides the leading order approximation of the MSE of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$. Generically we have $Q_\epsilon^{\mathrm{opt}}(\eta) = \Theta(n^{-1})$. Thus, as long as $\epsilon = o(n^{-2/3})$ the remainder term $O\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon^{\frac{1}{2}}\right)$ is asymptotically dominated by the leading term $Q_\epsilon^{\mathrm{opt}}(\eta)$.

Theorem D3 *Let Assumption D2 hold. Let $\widehat{\delta}_\epsilon = \widehat{\delta}_\epsilon(Y_1, \ldots, Y_n)$ be a sequence of estimators such that*

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\left[\widehat{\delta}_\epsilon - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^n h_\epsilon(Y_i,\eta)\right]^2 = o(\epsilon) + o(n^{-1}), \tag{D8}$$

for a sequence of influence functions $h_\epsilon(y,\eta)$ that satisfy (3) and (5). We then have, for $n \to \infty$ and $\epsilon \to 0$,

$$\sup_{\eta\in\mathcal{B}}\left\{\sup_{\theta_0\in\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \delta_{\theta_0}\right)^2\right] - \sup_{\theta_0\in\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0}\right)^2\right]\right\} = O\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon^{\frac{1}{2}}\right).$$

Remark: We could equivalently write $\widehat{\delta}_{n,\epsilon}$ and $h_{n,\epsilon}(y,\eta)$ in Theorem D3, that is, explicit dependence on n is allowed here. The theorem states that the worst-case MSE of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ is uniformly smaller than the worst-case MSE of any other estimators $\widehat{\delta}_\epsilon$ that satisfies suitable regularity conditions, up to remainder terms of order $O\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon^{\frac{1}{2}}\right)$.

Corollary D1 *Under the assumptions of Theorem D3 we have*

$$\begin{aligned}\int_{\mathcal{B}}\left\{\sup_{\theta_0\in\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_\epsilon^{\mathrm{MMSE}} - \delta_{\theta_0}\right)^2\right]\right\} w(\eta)\,d\eta \\ \leq \int_{\mathcal{B}}\left\{\sup_{\theta_0\in\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0}\right)^2\right]\right\} w(\eta)\,d\eta + O\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon^{\frac{1}{2}}\right),\end{aligned}$$

for any weight function $w : \mathcal{B} \to [0,\infty)$ such that the integrals in the last display are well-defined.

D.2 Proofs

For a given function $h = h(y,\eta)$ and parameter $\eta \in \mathcal{B}$ we define the $\dim\theta$ vector

$$v_{h,\eta} := \mathbb{E}_{\theta(\eta)}\left[h(Y,\eta)\,\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(Y)\right], \tag{D9}$$

and we also define

$$\begin{aligned}h^{\|}(y,\eta) &:= \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)\right]' \widetilde{H}^{\dagger}_{\theta(\eta)}\, v_{h,\eta} + \left[\nabla_\eta \log f_{\theta(\eta)}(y)\right]' H_\eta^{-1}\,\nabla_\eta \delta_{\theta(\eta)}, \\ h^{\perp}(y,\eta) &:= h(y,\eta) - h^{\|}(y,\eta).\end{aligned}$$

Lemma D2 *Let Assumption D2 be satisfied. Let $h = h(y,\eta)$ be such that the constraints* (3) *and* (5) *hold, and assume that $v_{h,\eta}$ defined in* (D9) *exist. Then, for $\epsilon > 0$ sufficiently small, and $(\theta_0, \eta) \in \Gamma_\epsilon$ we have*

(i) $\mathbb{E}_{\theta_0} h^{\parallel}(Y,\eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} = [\theta_0 - \theta(\eta)]' \left(v_{h,\eta} - \widetilde{\nabla}_\theta \delta_{\theta(\eta)} \right) + O\left(\epsilon + \epsilon \left\| v_{h,\eta} \right\|\right),$

(ii) $\mathbb{E}_{\theta_0} \left[h^{\parallel}(Y,\eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} \right]^2 = v_{h,\eta}' \widetilde{H}^{\dagger}_{\theta(\eta)} v_{h,\eta} + \left(\nabla_\eta \delta_{\theta(\eta)} \right)' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + O\left(\epsilon + \epsilon \left\| v_{h,\eta} \right\|^2\right),$

(iii) $\left| \mathbb{E}_{\theta_0} h^{\perp}(Y,\eta) \right| = O\left\{ \epsilon \left[\mathbb{E}_{\theta_0} h^{\perp}(Y,\eta)^2 \right]^{\frac{1}{2}} \right\},$

(iv) $\left| \mathbb{E}_{\theta_0} h^{\perp}(Y,\eta)\, h^{\parallel}(Y,\eta) \right| = O\left\{ \left(\epsilon^{\frac{3}{2}} + \epsilon^{\frac{1}{2}} \left\| v_{h,\eta} \right\| \right) \left[\mathbb{E}_{\theta_0} h^{\perp}(Y,\eta)^2 \right]^{\frac{1}{2}} \right\},$

where all the constants in the bounds $O(.)$ are independent of θ_0, η and $h(y,\eta)$.

Proof. # Part (i). Using the definition of $h^{\parallel}(y,\eta)$ we obtain

$$\begin{aligned}
&\mathbb{E}_{\theta_0} h^{\parallel}(Y,\eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} \\
&= \left[\mathbb{E}_{\theta_0} \nabla_\theta \log f_{\theta(\eta)}(Y) \right]' H^{\dagger}_{\theta(\eta)} v_{h,\eta} + \left[\mathbb{E}_{\theta_0} \nabla_\eta \log f_{\theta(\eta)}(Y) \right]' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} - \left[\delta_{\theta_0} - \delta_{\theta(\eta)} \right]
\end{aligned}$$

By mean value expansions in θ_0 around $\theta(\eta)$ we obtain

$$\begin{aligned}
\delta_{\theta_0} - \delta_{\theta(\eta)} &= [\theta_0 - \theta(\eta)]' \nabla_\theta \delta_{\theta(\eta)} + O(\epsilon), \\
\mathbb{E}_{\theta_0} \widetilde{\nabla}_\theta \log f_{\theta(\eta)}(Y) &= \mathbb{E}_{\theta(\eta)} \left[\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(Y) \right] \left[\nabla_\theta \log f_{\theta(\eta)}(Y) \right]' [\theta_0 - \theta(\eta)] + O(\epsilon) \\
&= \widetilde{H}_{\theta(\eta)} [\theta_0 - \theta(\eta)] + O(\epsilon), \\
\mathbb{E}_{\theta_0} \nabla_\eta \log f_{\theta(\eta)}(Y) &= G_\eta\, \mathbb{E}_{\theta(\eta)} \left[\nabla_\theta \log f_{\theta(\eta)}(Y) \right] \left[\nabla_\theta \log f_{\theta(\eta)}(Y) \right]' [\theta_0 - \theta(\eta)] + O(\epsilon) \\
&= G_\eta\, H_{\theta(\eta)} [\theta_0 - \theta(\eta)] + O(\epsilon).
\end{aligned}$$

We have $\widetilde{H}_{\theta(\eta)} \widetilde{H}^{\dagger}_{\theta(\eta)} v_{h,\eta} = v_{h,\eta}$, because the column span of $H_{\theta(\eta)}$ is equal to the linear span of $\widetilde{\nabla}_\theta \log f_{\theta(\eta)}(y)$ over all values y with $f_{\theta(\eta)}(y) > 0$, and $v_{h,\eta}$ lies in that linear span. Combing the above gives

$$\mathbb{E}_{\theta_0} h^{\parallel}(Y,\eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} = [\theta_0 - \theta(\eta)]' \left[v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)} + H_{\theta(\eta)}\, G_\eta' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} \right] + O\left(\epsilon + \epsilon \left\| v_{h,\eta} \right\|\right),$$

and using the definition of $\widetilde{\nabla}_\theta$ this gives the the statement of part (i) of the lemma.

Part (ii). By defining

$$v^*_{h,\eta} := v_{h,\eta} + H_{\theta(\eta)} G_\eta' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)},$$

we can rewrite $h^{\|}(y,\eta)$ as

$$h^{\|}(y,\eta) = \left[\nabla_\theta \log f_{\theta(\eta)}(y)\right]' H^{\dagger}_{\theta(\eta)} v^*_{h,\eta}.$$

Using this we obtain

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[h^{\|}(Y,\eta) + \delta_{\theta(\eta)} - \delta_{\theta_0}\right]^2 = \left[\delta_{\theta_0} - \delta_{\theta(\eta)}\right]^2 - 2\left[\delta_{\theta_0} - \delta_{\theta(\eta)}\right]\left[\mathbb{E}_{\theta_0}\nabla_\theta \log f_{\theta(\eta)}(Y)\right]' H^{\dagger}_{\theta(\eta)} v^*_{h,\eta} \\ + v^{*\prime}_{h,\eta} H^{\dagger}_{\theta(\eta)} \mathbb{E}_{\theta_0}\left\{\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]'\right\} H^{\dagger}_{\theta(\eta)} v^*_{h,\eta}.\end{aligned}$$

By mean value expansions in θ_0 around $\theta(\eta)$ we obtain

$$\begin{aligned}\delta_{\theta_0} - \delta_{\theta(\eta)} &= O(\epsilon^{\frac{1}{2}}), \\ \mathbb{E}_{\theta_0}\nabla_\theta \log f_{\theta(\eta)}(Y) &= O(\epsilon^{\frac{1}{2}}), \\ \mathbb{E}_{\theta_0}\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]' &= \mathbb{E}_{\theta(\eta)}\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]\left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]' + O(\epsilon^{\frac{1}{2}}) \\ &= H_{\theta(\eta)} + O(\epsilon^{\frac{1}{2}}).\end{aligned}$$

Combining the above, and using $H_{\theta(\eta)}H^{\dagger}_{\theta(\eta)}v^*_{h,\eta} = v^*_{h,\eta}$, gives

$$\mathbb{E}_{\theta_0}\left[h^{\|}(Y,\eta) + \delta_{\theta(\eta)} - \delta_{\theta_0}\right]^2 = v^{*\prime}_{h,\eta} H^{\dagger}_{\theta(\eta)} v^*_{h,\eta} + O\left(\epsilon + \epsilon \left\|v^*_{h,\eta}\right\|^2\right),$$

and by plugging in the definition of $v^*_{h,\eta}$ and using that $G_\eta v^*_{h,\eta} = 0$ we obtain the statement of part (ii) of the lemma.

Part (iii). The constraints (3) and (5) guarantee that $\mathbb{E}_{\theta(\eta)}h^{\perp}(Y,\eta) = 0$ as well as $\mathbb{E}_{\theta(\eta)}\left[h^{\perp}(Y,\eta)\,\nabla_\theta \log f_{\theta(\eta)}(Y)\right] = 0$. Using this we find

$$\begin{aligned}\mathbb{E}_{\theta_0}h^{\perp}(Y,\eta) &= \mathbb{E}_{\theta_0}h^{\perp}(Y,\eta) - \mathbb{E}_{\theta(\eta)}h^{\perp}(Y,\eta) - [\theta_0 - \theta(\eta)]'\,\mathbb{E}_{\theta(\eta)}\left[h^{\perp}(Y,\eta)\,\nabla_\theta \log f_{\theta(\eta)}(Y)\right] \\ &= \int_{\mathcal{Y}} h^{\perp}(y,\eta)\,\left\{f_{\theta_0}(y) - f_{\theta(\eta)}(y) - [\theta_0 - \theta(\eta)]'\nabla_\theta f_{\theta(\eta)}(y)\right\} dy. \\ &= \int_{\mathcal{Y}} h^{\perp}(y,\eta)\sqrt{f_{\theta_0}(y)}\,\frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y) - [\theta_0 - \theta(\eta)]'\nabla_\theta f_{\theta(\eta)}(y)}{\sqrt{f_{\theta_0}(y)}}\,dy.\end{aligned}$$

Applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}\left[\mathbb{E}_{\theta_0}h^{\perp}(Y,\eta)\right]^2 &\le \int_{\mathcal{Y}} h^{\perp}(y,\eta)^2 f_{\theta_0}(y)dy \int_{\mathcal{Y}} \frac{\left\{f_{\theta_0}(y) - f_{\theta(\eta)}(y) - [\theta_0 - \theta(\eta)]'\nabla_\theta f_{\theta(\eta)}(y)\right\}^2}{f_{\theta_0}(y)}dy \\ &= \left[\mathbb{E}_{\theta_0}h^{\perp}(Y,\eta)^2\right] \underbrace{\mathbb{E}_{\theta_0}\left(1 - \frac{f_{\theta(\eta)}(Y) + [\theta_0 - \theta(\eta)]'\nabla_\theta f_{\theta(\eta)}(Y)}{f_{\theta_0}(Y)}\right)^2}_{:= g_{\theta_0}}.\end{aligned}$$

The function g_{θ_0} satifies $g_{\theta(\eta)} = 0$, $\nabla_\theta g_{\theta(\eta)} = 0$, $\nabla_{\theta\theta'} g_{\theta(\eta)} = 0$, $\nabla_{\theta_k\theta_\ell\theta_m} g_{\theta(\eta)} = 0$, for all $k, \ell, m \in \{1, \ldots, \dim\theta\}$. By a mean value expansion in θ_0 around $\theta(\eta)$ we therefore obtain that $g_{\theta_0} = O(\epsilon^2)$. Combing the above gives the statement of part (iii) of the lemma.

Part (iv). We have

$$\mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta_0}\right] = \mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta(\eta)}\right] + \left(\delta_{\theta(\eta)} - \delta_{\theta_0}\right) \mathbb{E}_{\theta_0} h^\perp(Y,\eta),$$

and therefore

$$\begin{aligned}\left|\mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta_0}\right]\right| &\leq \left|\mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta(\eta)}\right]\right| + \left|\delta_{\theta(\eta)} - \delta_{\theta_0}\right| \left|\mathbb{E}_{\theta_0} h^\perp(Y,\eta)\right| \\ &= \left|\mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta(\eta)}\right]\right| + O\left\{\epsilon^{\frac{3}{2}} \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)^2\right]^{\frac{1}{2}}\right\},\end{aligned}$$

where in the last step we used result on $\delta_{\theta(\eta)} - \delta_{\theta_0}$ and $\mathbb{E}_{\theta_0} h^\perp(Y,\eta)$ already shown above. Next, because $\mathbb{E}_{\theta(\eta)} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta(\eta)}\right] = 0$ we have

$$\begin{aligned}&\mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta(\eta)}\right] = \int_{\mathcal{Y}} h^\perp(y,\eta) \left[h^\parallel(y,\eta) - \delta_{\theta(\eta)}\right] \left[f_{\theta_0}(y) - f_{\theta(\eta)}(y)\right] dy \\ &= \int_{\mathcal{Y}} h^\perp(y,\eta) \sqrt{f_{\theta_0}(y)} \left[h^\parallel(Y,\eta) - \delta_{\theta(\eta)}\right] \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{\sqrt{f_{\theta_0}(y)}} dy.\end{aligned}$$

Applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}&\left\{\mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \delta_{\theta_0}\right]\right\}^2 \\ &\leq \int_{\mathcal{Y}} h^\perp(y,\eta)^2 f_{\theta_0}(y) dy \int_{\mathcal{Y}} \left[h^\parallel(y,\eta) - \delta_{\theta(\eta)}\right]^2 \frac{\left[f_{\theta_0}(y) - f_{\theta(\eta)}(y)\right]^2}{f_{\theta_0}(y)} dy \\ &= \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)^2\right] v'_{h,\eta} H^\dagger_{\theta(\eta)} \underbrace{\left[\mathbb{E}_{\theta_0} \left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right] \left[\nabla_\theta \log f_{\theta(\eta)}(Y)\right]' \left(\frac{f_{\theta(\eta)}(Y)}{f_{\theta_0}(Y)} - 1\right)^2\right]}_{=: G_{\theta_0}} H^\dagger_{\theta(\eta)} v_{h,\eta}.\end{aligned}$$

The matrix valued function G_{θ_0} satisfies $G_{\theta(\eta)} = 0$ and $\nabla_{\theta_k} G_{\theta(\eta)} = 0$. By a mean value expansion in θ_0 around $\theta(\eta)$ we therefore obtain that $G_{\theta(\eta)} = O(\epsilon)$. Combing the above gives the statement of part (iv) of the lemma. ■

For a function $h = h(y,\eta)$ let

$$Q_\epsilon(h,\eta) = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} Q^*_\epsilon(h,\eta,\theta_0), \qquad Q^*_\epsilon(h,\eta,\theta_0) = \mathbb{E}_{\theta_0} \left(\frac{1}{n}\sum_{i=1}^n h(Y_i,\eta) + \delta_{\theta(\eta)} - \delta_{\theta_0}\right)^2.$$

Lemma D3 *Let Assumption D2 be satisfied. Let $\eta \in \mathcal{B}$. Let $h = h(y,\eta)$ be such that the constraints* (3) *and* (5) *hold. Let $v_{h,\eta}$ be as defined in* (D9)*. Assume that $Q_\epsilon(h,\eta)$ exists, which also guarantees existence of $v_{h,\eta}$. We then have*

(i) $Q_\epsilon(h,\eta) \geq Q_\epsilon^{\text{low}}\left(v_{h,\eta}, \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)^2\right]^{\frac{1}{2}}, \eta\right)$, *where*

$$Q_\epsilon^{\text{low}}(v,s,\eta) := \epsilon \left\| v - \widetilde{\nabla}_\theta \delta_{\theta(\eta)} \right\|_\eta^2 + \frac{1}{n} v' \widetilde{H}^\dagger_{\theta(\eta)} v + \frac{1}{n} \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + \frac{1}{n} s^2 - 2\, c_\eta \Big[(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon)(1 + \|v\|^2) + \left(\epsilon^2 + n^{-1}\epsilon^{\frac{3}{2}} + \epsilon^2 \|v\| + n^{-1}\epsilon^{\frac{1}{2}} \|v\|\right) s \Big],,$$

for some non-random $c_\eta > 0$ *with* $\sup_{\eta \in \mathcal{B}} c_\eta < \infty$.

(ii) If also $h^\perp(y,\eta) = 0$, *then*

$$Q_\epsilon(h,\eta) = \epsilon \left\| v_{h,\eta} - \widetilde{\nabla}_\theta \delta_{\theta(\eta)} \right\|_\eta^2 + \frac{1}{n} v_{h,\eta}' \widetilde{H}^\dagger_{\theta(\eta)} v_{h,\eta} + \frac{1}{n} \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + O\left[(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon)(1 + \|v_{h,\eta}\|^2) \right].$$

Proof. Let $\Delta_{\eta,\theta_0} = \delta_{\theta_0} - \delta_{\theta(\eta)}$. We have

$$\begin{aligned}
&Q_\epsilon^*(h,\eta,\theta_0) \\
&= \left[\mathbb{E}_{\theta_0} h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \frac{1}{n} \operatorname{Var}_{\theta_0}\left[h(Y,\eta) - \Delta_{\eta,\theta_0}\right] \\
&= \frac{n-1}{n} \left[\mathbb{E}_{\theta_0} h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \frac{1}{n} \mathbb{E}_{\theta_0}\left[h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 \\
&= \frac{n-1}{n} \left[\mathbb{E}_{\theta_0} h^\parallel(Y,\eta) - \Delta_{\eta,\theta_0} + \mathbb{E}_{\theta_0} h^\perp(Y,\eta)\right]^2 + \frac{1}{n} \mathbb{E}_{\theta_0} \left[\mathbb{E}_{\theta_0} h^\parallel(Y,\eta) - \Delta_{\eta,\theta_0} + \mathbb{E}_{\theta_0} h^\perp(Y,\eta)\right]^2 \\
&= \frac{n-1}{n} \left\{ \left[\mathbb{E}_{\theta_0} h^\parallel(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)\right]^2 + 2 \left[\mathbb{E}_{\theta_0} h^\parallel(Y,\eta) - \Delta_{\eta,\theta_0}\right] \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)\right] \right\} \\
&\quad + \frac{1}{n} \left\{ \mathbb{E}_{\theta_0} \left[h^\parallel(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \mathbb{E}_{\theta_0} h^\perp(Y,\eta)^2 + 2\, \mathbb{E}_{\theta_0} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \Delta_{\eta,\theta_0}\right] \right\}. \quad \text{(D10)}
\end{aligned}$$

Let[29]

$$\theta^* \in \operatorname*{argmax}_{\theta_0 \in \Gamma_\epsilon(\eta)} \left[\theta_0 - \theta(\eta)\right]' \left(v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)}\right).$$

Notice that θ^* depends on $h = h(y,\eta)$ and η and ϵ, but we suppress this dependence here for notational convenience. We have

$$\begin{aligned}
Q_\epsilon(h,\eta) &\geq Q_\epsilon^*(h,\eta,\theta^*) \\
&\geq \frac{n-1}{n} \left\{ \left[\mathbb{E}_{\theta^*} h^\parallel(Y,\eta) - \Delta_{\eta,\theta^*}\right]^2 - 2 \left|\mathbb{E}_{\theta^*} h^\parallel(Y,\eta) - \Delta_{\eta,\theta^*}\right| \left|\mathbb{E}_{\theta^*} h^\perp(Y,\eta)\right| \right\} \\
&\quad + \frac{1}{n} \left\{ \mathbb{E}_{\theta^*} \left[h^\parallel(Y,\eta) - \Delta_{\eta,\theta^*}\right]^2 + \mathbb{E}_{\theta^*} h^\perp(Y,\eta)^2 - 2 \left|\mathbb{E}_{\theta^*} h^\perp(Y,\eta) \left[h^\parallel(Y,\eta) - \Delta_{\eta,\theta^*}\right]\right| \right\}.
\end{aligned}$$

[29]This a maximization of a continuous function of θ_0 over a compact set, so existence of a maximizing value θ^* is guaranteed, but it need not be unique.

Using Lemma D2 we obtain

$$\begin{aligned}
\mathbb{E}_{\theta^*} h^{\|}(Y,\eta) - \Delta_{\eta,\theta^*} &= [\theta^* - \theta(\eta)]' \left(v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)}\right) + O(\epsilon) + O\left(\epsilon \left\|v_{h,\eta}\right\|\right) \\
&= \epsilon^{\frac{1}{2}} \left\|v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)}\right\|_\eta + O(\epsilon) + O\left(\epsilon \left\|v_{h,\eta}\right\|\right), \\
\left|\mathbb{E}_{\theta^*} h^{\|}(Y,\eta) - \Delta_{\eta,\theta^*}\right| &= O\left(\epsilon^{\frac{1}{2}}\right) + O\left(\epsilon^{\frac{1}{2}} \left\|v_{h,\eta}\right\|\right), \\
\mathbb{E}_{\theta^*}\left[h^{\|}(Y,\eta) - \Delta_{\eta,\theta^*}\right]^2 &= v'_{h,\eta} H^\dagger_{\theta(\eta)} v_{h,\eta} + O(\epsilon) + O\left(\epsilon \left\|v_{h,\eta}\right\|^2\right), \\
\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta) &= O\left\{\epsilon \left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\}, \\
\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)\left[h^{\|}(Y,\eta) - \Delta_{\eta,\theta^*}\right] &= O\left\{\left(\epsilon^{\frac{3}{2}} + \epsilon^{\frac{1}{2}} \left\|v_{h,\eta}\right\|\right)\left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\},
\end{aligned}$$

and therefore

$$\begin{aligned}
Q_\epsilon(h,\eta) \geq{}& \epsilon \left\|v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)}\right\|_\eta^2 + O\left(\epsilon^{\frac{3}{2}}\right) + O\left(\epsilon^{\frac{3}{2}} \left\|v_{h,\eta}\right\|^2\right) + O\left(n^{-1}\epsilon\right) + O\left(n^{-1}\epsilon \left\|v_{h,\eta}\right\|^2\right) \\
&+ \left[O(\epsilon) + O\left(\epsilon \left\|v_{h,\eta}\right\|\right)\right] O\left\{\epsilon \left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\} \\
&+ \frac{1}{n}\left\{v'_{h,\eta} H^\dagger_{\theta(\eta)} v_{h,\eta} + O(\epsilon) + O\left(\epsilon \left\|v_{h,\eta}\right\|^2\right) + \mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right\} \\
&+ O\left\{n^{-1}\left(\epsilon^{\frac{3}{2}} + \epsilon^{\frac{1}{2}} \left\|v_{h,\eta}\right\|\right)\left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\} \\
={}& \epsilon \left\|v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)}\right\|_\eta^2 + \frac{1}{n} v'_{h,\eta} H^\dagger_{\theta(\eta)} v_{h,\eta} + \frac{1}{n}\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2 \\
&+ O\left(\epsilon^{\frac{3}{2}}\right) + O\left(\epsilon^{\frac{3}{2}} \left\|v_{h,\eta}\right\|^2\right) + O\left(n^{-1}\epsilon\right) + O\left(n^{-1}\epsilon \left\|v_{h,\eta}\right\|^2\right) \\
&+ O\left\{\epsilon^2 \left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\} + O\left\{\epsilon^2 \left\|v_{h,\eta}\right\| \left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\} \\
&+ O\left(n^{-1}\epsilon\right) + O\left(n^{-1}\epsilon \left\|v_{h,\eta}\right\|^2\right) \\
&+ O\left\{n^{-1}\left(\epsilon^{\frac{3}{2}} + \epsilon^{\frac{1}{2}} \left\|v_{h,\eta}\right\|\right)\left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\} \\
={}& \epsilon \left\|v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)}\right\|_\eta^2 + \frac{1}{n} v'_{h,\eta} H^\dagger_{\theta(\eta)} v_{h,\eta} + \frac{1}{n}\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2 \\
&+ O\left(\epsilon^{\frac{3}{2}}\right) + O\left(\epsilon^{\frac{3}{2}} \left\|v_{h,\eta}\right\|^2\right) + O\left(n^{-1}\epsilon\right) + O\left(n^{-1}\epsilon \left\|v_{h,\eta}\right\|^2\right) \\
&+ O\left\{\left(\epsilon^2 + n^{-1}\epsilon^{\frac{3}{2}} + \epsilon^2 \left\|v_{h,\eta}\right\| + n^{-1}\epsilon^{\frac{1}{2}} \left\|v_{h,\eta}\right\|\right)\left[\mathbb{E}_{\theta^*} h^{\perp}(Y,\eta)^2\right]^{\frac{1}{2}}\right\}
\end{aligned}$$

This is the lower bound on $Q_\epsilon(h,\eta)$ given in part (i) of the lemma.

Next, if $h^{\perp}(y,\eta) = 0$, then equation (D10) simplifies to

$$\begin{aligned}
Q^*_\epsilon(h,\theta_0) &= \frac{n-1}{n}\left[\mathbb{E}_{\theta_0} h^{\|}(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \frac{1}{n}\mathbb{E}_{\theta_0}\left[h^{\|}(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 \\
&= \frac{n-1}{n}\left\{[\theta_0 - \theta(\eta)]'\left(v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)}\right) + O(\epsilon) + O\left(\epsilon \left\|v_{h,\eta}\right\|\right)\right\}^2 \\
&\quad + \frac{1}{n}\left\{v'_{h,\eta} H^\dagger_{\theta(\eta)} v_{h,\eta} + O(\epsilon) + O\left(\epsilon \left\|v_{h,\eta}\right\|^2\right)\right\},
\end{aligned}$$

where in the last step we used Lemma D2. Taking the supremum over θ_0, and using

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} [\theta_0 - \theta(\eta)]' \left(v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)} \right) = \epsilon^{\frac{1}{2}} \left\| v_{h,\eta} - \nabla_\theta \delta_{\theta(\eta)} \right\|_\eta$$

gives the statement of part (ii) of the lemma. ■

Lemma D4 *Let Assumption D2 hold. Let $Q_\epsilon^{\text{opt}}(\eta)$ be as defined before Theorem D2. Let $\widehat{\delta}_\epsilon = \widehat{\delta}_\epsilon(Y_1, \ldots, Y_n)$ satisfy the assumptions of Theorem D3. We then have, for $n \to \infty$ and $\epsilon \to 0$,*

$$\sup_{\eta \in \mathcal{B}} \left\{ Q_\epsilon^{\text{opt}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0} \right)^2 \right] \right\} \leq O\left(\epsilon^{\frac{3}{2}} + n^{-1} \epsilon^{\frac{1}{2}} \right).$$

Proof. Using (D8) and the definition of $Q_\epsilon(h, \eta)$ we find that

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

Thus, what is left to show is that

$$\sup_{\eta \in \mathcal{B}} \left[Q_\epsilon^{\text{opt}}(\eta) - Q_\epsilon(h_\epsilon, \eta) \right] \leq O\left(\epsilon^{\frac{3}{2}} + n^{-1} \epsilon^{\frac{1}{2}} \right). \tag{D11}$$

In the following we just write $h = h(y, \eta)$ instead of $h_\epsilon = h_\epsilon(y, \eta)$.

We have $\sup_{\eta \in \mathcal{B}} Q_\epsilon^{\text{opt}}(\eta) = O(n^{-1})$. Thus, if $n^{-1} = O(\epsilon^{\frac{3}{2}})$, then we have $\sup_{\eta \in \mathcal{B}} Q_\epsilon^{\text{opt}}(\eta) = O(\epsilon^{\frac{3}{2}})$, and the statement in the lemma holds trivially. For the remainder of this proof we can therefore consider the case

$$n\, \epsilon^{\frac{3}{2}} \to 0. \tag{D12}$$

According to Lemma D3 we then have $Q_\epsilon(h,\eta) \geq Q_\epsilon^{\text{low}}\left(v_{h,\eta}, \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)^2\right]^{\frac{1}{2}}, \eta\right)$. By solving the quadratic minimization over s we obtain

$$
\begin{aligned}
&Q_\epsilon^{\text{low}}\left(v, \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)^2\right]^{\frac{1}{2}}\right) \\
&\quad \geq \min_{s\geq 0} Q_\epsilon^{\text{low}}(v,s,\eta) \\
&\quad = \epsilon \left\| v - \widetilde{\nabla}_\theta \delta_{\theta(\eta)} \right\|_\eta^2 + \frac{1}{n}\, v'\, \widetilde{H}^\dagger_{\theta(\eta)}\, v + \frac{1}{n}\, \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1}\, \nabla_\eta \delta_{\theta(\eta)} \\
&\qquad - 2\, c_\eta \left[\epsilon^{\frac{3}{2}} + n^{-1}\epsilon + \epsilon^{\frac{3}{2}} \|v\|^2 + n^{-1}\epsilon \|v\|^2\right] \\
&\qquad - c_\eta^2\, n \left(\epsilon^2 + n^{-1}\epsilon^{\frac{3}{2}} + \epsilon^2 \|v\| + n^{-1}\epsilon^{\frac{1}{2}} \|v\|\right)^2 \\
&\quad \geq \epsilon \left\| v - \widetilde{\nabla}_\theta \delta_{\theta(\eta)} \right\|_\eta^2 + \frac{1}{n}\, v'\, \widetilde{H}^\dagger_{\theta(\eta)}\, v + \frac{1}{n}\, \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1}\, \nabla_\eta \delta_{\theta(\eta)} \\
&\qquad\qquad + O\left[\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon + n\,\epsilon^4\right)\left(1 + \|v\|^2\right)\right] \\
&\quad = \epsilon \left\| v - \widetilde{\nabla}_\theta \delta_{\theta(\eta)} \right\|_\eta^2 + \frac{1}{n}\, v'\, \widetilde{H}^\dagger_{\theta(\eta)}\, v + \frac{1}{n}\, \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1}\, \nabla_\eta \delta_{\theta(\eta)} \\
&\qquad\qquad + O\left[\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon\right)\left(1 + \|v\|^2\right)\right], \qquad \text{(D13)}
\end{aligned}
$$

where in the last step we used that $n\,\epsilon^4 = O(\epsilon^{\frac{3}{2}})$ under (D12). The condition (D12) also guarantees that as $n \to \infty$ and $\epsilon \to 0$ we have

$$\min\left(\epsilon, \frac{1}{n}\right) \gg \epsilon^{\frac{3}{2}} + n^{-1}\epsilon.$$

Thus, when minimizing the right hand side of (D13) over $v \in \mathbb{R}^{\dim\theta}$, then the terms $\left\| v - \widetilde{\nabla}_\theta \delta_{\theta(\eta)} \right\|_\eta^2 + \frac{1}{n}\, v'\, \widetilde{H}^\dagger_{\theta(\eta)}\, v$ will dominate and the minimizing value for v will thus satisfy $v^* = O(1)$. We therefore find

$$
\begin{aligned}
Q_\epsilon(h,\eta) &\geq \min_{v\in\mathbb{R}^{\dim\theta}} Q_\epsilon^{\text{low}}\left(v, \left[\mathbb{E}_{\theta_0} h^\perp(Y,\eta)^2\right]^{\frac{1}{2}}, \eta\right) \\
&\geq Q_\epsilon^{\text{opt}}(\eta) + O\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon\right),
\end{aligned}
$$

where in the last step we used that $n\,\epsilon^4 = O(\epsilon^{\frac{3}{2}})$ under (D12). All constants hidden $O(.)$ above are uniformly bounded over η, and we thus have (D11). ■

Proof of Theorem D2. It is easy to see that $\widehat{\delta}_\epsilon^{\text{MMSE}}$ satisfies the regularity conditions for general $\widehat{\delta}_\epsilon$ in Theorem D3. Applying Lemma D4 we thus have

$$\sup_{\eta\in\mathcal{B}} \left\{ Q_\epsilon^{\text{opt}}(\eta) - \sup_{\theta_0\in\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\theta_0}\right)^2\right] \right\} \leq O\left(\epsilon^{\frac{3}{2}} + n^{-1}\epsilon^{\frac{1}{2}}\right).$$

We also have

$$\begin{aligned}
&\sup_{\eta\in\mathcal{B}}\left\{\sup_{\theta_0\in\Gamma_\epsilon(\eta)}\mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}}-\delta_{\theta_0}\right)^2\right]-Q_\epsilon^{\text{opt}}(\eta)\right\}\\
&=\sup_{\eta\in\mathcal{B}}\left\{\sup_{\theta_0\in\Gamma_\epsilon(\eta)}\mathbb{E}_{\theta_0}\left[\left(\frac{1}{n}\sum_{i=1}^n h_\epsilon^{\text{MMSE}}(Y_i,\eta)+\delta_{\theta(\eta)}-\delta_{\theta_0}\right)^2\right]-Q_\epsilon^{\text{opt}}(\eta)\right\}\\
&=\sup_{\eta\in\mathcal{B}}\left[Q_\epsilon(h_\epsilon^{\text{MMSE}},\eta)-Q_\epsilon^{\text{opt}}(\eta)\right]\\
&\leq O\left(\epsilon^{\frac{3}{2}}+n^{-1}\epsilon^{\frac{1}{2}}\right),
\end{aligned}$$

where the last step follows from part (ii) Lemma D3. ■

Proof of Theorem D3. Now follows immediately from Theorem D2 and Lemma D4. ■

E Two additional semi-parametric examples

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

E.1 A demand model

In our first 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,

$$Y_i = j \;\Leftrightarrow\; X_{ij}'\beta_j + A_{ij} \geq X_{ik}'\beta_k + A_{ik} \text{ for all } k \neq j. \tag{E14}$$

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 $\delta_{\theta_0} = \mathbb{E}_{\theta_0}(\Delta(A, X, \beta_0))$, 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, De Palma and Thisse (1992), for example.

Assuming that β and γ are known for simplicity, in this example we have, by (34) and (36),

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

where

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

and, for all $k = 1, ..., K$ and x,

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

E.2 Average treatment effects under selection on observables

In our second 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

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

and

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

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

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

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

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

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.

F Bayesian interpretation

F.1 Gaussian prior

Consider the known η case to start with. To see that (49) holds, note that, under sufficient smoothness,

$$\mathbb{E}\left[\delta_{\theta_0} \,|\, Y_1,\ldots,Y_n,\eta\right] = \delta_{\theta(\eta)} + (\nabla_\theta \delta_{\theta(\eta)})'\mathbb{E}\left[\theta_0-\theta(\eta) \,|\, Y_1,\ldots,Y_n,\eta\right] + o_P(\epsilon^{\frac{1}{2}}), \tag{F15}$$

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 independent of θ_0. Under mild regularity conditions, taking expectations in (F15) with respect to the posterior distribution of η implies that (49) holds.

F.2 Least favorable prior

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

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

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

When the order of the min and max can be reversed, the least-favorable prior ρ^{LF} solves

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

For given h the integral is equal to

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

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

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

that is,

$$h^*(y,\eta) = \delta_{\theta(\eta)} + n\,\nabla_\theta \log f_{\theta(\eta)}(y)' \left[n\,H_{\theta(\eta)} + \Omega^{\frac{1}{2}} V_\Omega(\rho)^{-1}\Omega^{\frac{1}{2}}\right]^{-1} \nabla_\theta \delta_{\theta(\eta)}.$$

Moreover, it is easy to show that

$$\int_{\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[(\widehat{\delta}_{h^*,\eta} - \delta_{\theta_0})^2\right] \rho(\theta_0)d\theta_0 = (\nabla_\theta \delta_{\theta(\eta)})' \left[n\,H_{\theta(\eta)} + \Omega^{\frac{1}{2}} V_\Omega(\rho)^{-1}\Omega^{\frac{1}{2}}\right]^{-1} \nabla_\theta \delta_{\theta(\eta)}.$$

Now, $\epsilon^{-1}V_\Omega(\rho)$ is bounded by the identity matrix. Any ρ satisfying

$$V_\Omega(\rho^{\mathrm{LF}}) = \epsilon\, I_{\dim\theta}$$

is thus minimax, and ρ^{LF} puts all mass at the boundary of $\Gamma_\epsilon(\eta)$. It is easy to see that the optimal h^*, for ρ^{LF}, coincides with $h_\epsilon^{\mathrm{MMSE}}$.

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

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

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 (16).

Applying the above arguments, one sees that the least-favorable prior satisfies

$$V_\Omega(\rho^{\mathrm{LF}}(\cdot\,|\,\eta)) = \epsilon\, I_{\dim\theta}, \text{ for all } \eta.$$

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

G Extensions

G.1 Fixed-ϵ bias

The derivations in this subsection are heuristic. 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

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

The first-order conditions with respect to π_0 are then

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

Hence, using that π_0 integrates to one,

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

where

$$C^{-1} = \int_{\mathcal{A}} \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x)g(y \mid a)dy\right)\right)\pi(a)da. \tag{G16}$$

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

$$\epsilon = 2\log C - \frac{C}{\lambda_2}\int_{\mathcal{A}} \left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x)g(y \mid a)dy\right) \times \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x)g(y \mid a)dy\right)\right)\pi(a)da. \tag{G17}$$

It follows that

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

where C and λ_2 satisfy (G16)-(G17).

Hence (50) follows.

G.2 Local equivalence to worst-case bounds

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

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

This shows (51).

To see when (51) 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

$$\sup_{(\theta_1,\theta_2)\in\Gamma_\epsilon(\eta)^2} \delta_{\theta_2} - \delta_{\theta_1} + \int_{\mathcal{Y}} \lambda_1(y) f_{\theta_1}(y) dy + \int_{\mathcal{Y}} \lambda_2(y) f_{\theta_2}(y) dy, \tag{G18}$$

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, (G18) is equal to

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

where we have used (52).

G.3 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

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

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,

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

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

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

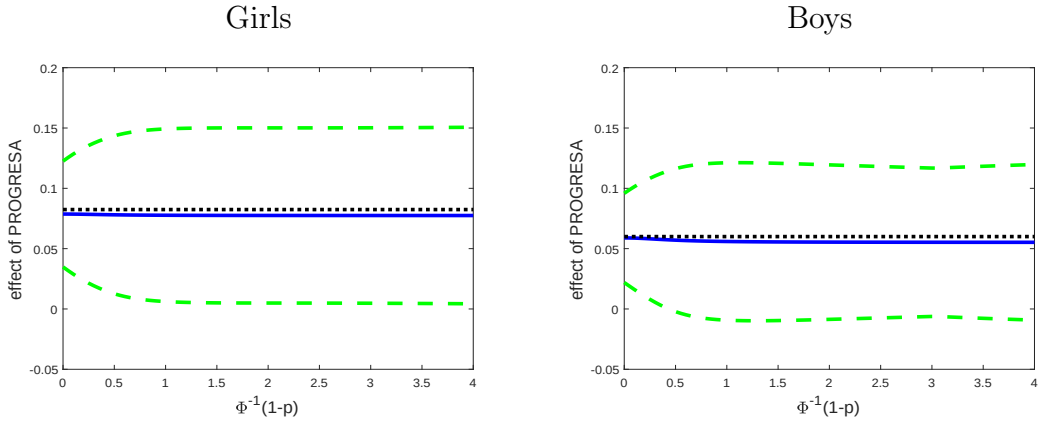
H Conditional cash transfers in Mexico, reference model estimated on both control and treated villages

Table H1: 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 H1: 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 (29), 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).