Minimizing Sensitivity to Model Misspecification*

Stéphane Bonhomme† Martin Weidner‡

January 2018

PRELIMINARY AND INCOMPLETE DRAFT

Abstract

We propose a framework to compute predictions based on an economic model when the model may be misspecified. Our approach relies on minimizing sensitivity of the estimates to the type of misspecification that is most influential for the parameter of interest. We rely on a local asymptotic approach where the degree of misspecification is indexed by the sample size. This results in simple rules to adjust the predictions from the reference model. We calibrate the degree of misspecification using a detection error probability approach, which allows us to perform systematic sensitivity analysis in both point-identified and partially-identified settings. We study three examples: demand analysis, treatment effects estimation under selection on observables, and panel data models where the distribution of individual effects may be misspecified and the number of time periods is small.

KEYWORDS: Model misspecification, robustness, sensitivity, latent variables, panel data.

*We thank Kei Hirano, Thibaut Lamadon, Alex Torgovistky, and Whitney Newey for comments. Bonhomme acknowledges support from the NSF, Grant SES-1658920.

†University of Chicago.

‡University College London.

1 Introduction

Although economic models are intended to be 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 models to compute model-implied predictions, such as counterfactual effects of policies. Given previously estimated model parameters, it is common practice to simply "plug in" those parameters in a formula giving the effect of interest. Such a practice, which typically requires a full specification of the economic environment, hinges on the model to be correctly specified.

Economists have long recognized the dangers of risk of misspecification. Traditional reactions include specification tests and estimation of more general nesting models, semi- and nonparametric modeling, and more recently bounds approaches. In particular, those approaches have proven useful to relax functional form assumptions that are rarely motivated by economic arguments. However, they require 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 its impact. The goal of the analysis is to provide simple adjustments of the model-based quantities, which do not require re-estimating the model and provide guarantees on performance when the model is misspecified.

Our approach is based on considering deviations away from a *reference specification* of the model. This reference 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 from it. While it is theoretically possible to extend our approach to allow for any types of deviations (and we also provide a *global* characterization of misspecification bias below), a local analysis presents important advantages in terms of tractability since it allows us to rely on linearization techniques.

We adopt a *minimax* approach, in the sense that our aim is to construct estimators which minimize the worst-case bias or mean-squared error in a given neighborhood of the reference model. This worst case is influenced by the directions of model misspecification which matter most for the target quantity of interest. The neighborhoods we consider are in terms of a locally quadratic metric. A leading example of the latter is the Kullback-Leibler

divergence, which we use in order to measure discrepancies between the reference model and alternative models that may have generated the data.

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. This leads to a class of estimates 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 this approach, acknowledging misspecification leads to easy-to-compute enlargements of conventional confidence intervals.

Our local approach leads to tractable expressions for bias and mean squared error as well as their minimizers. In particular, minimum-mean squared error estimators generically take the form of an adjustment of the prediction based on the reference model by a term which reflects the impact of misspecification. The latter is scaled by the product of the size of the neighborhood and the sample size. All quantities can readily be computed by simulation given the reference model.

We study three illustrations. The first one is a model of demand for differentiated products. We consider a setting where the researcher has estimated a reference model, such as a multinomial logit model, and wishes to use the model to compute a counterfactual market share or an elasticity to some input. The concern for misspecification of the logistic assumption leads to adjusting the logit-based formula by a term which minimizes the impact of model misspecification on the particular target parameter the researcher wishes to compute.

Our second illustration concerns the estimation of an average treatment effect under selection on observables. In a model with known propensity score, we show that minimizing sensitivity with respect to the conditional mean of potential outcomes leads to inverse propensity weighting estimators. In turn, minimum-mean squared error estimators are weighted averages of the inverse propensity weights and a model-based expression for the mean potential outcomes.

Our third and main illustration is a class of panel data models which covers both static and dynamic settings. We focus on average effects, which depend on the distribution of individual effects. The risk of misspecification of this distribution and its dependence on co-

variates 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), we focus on assessing sensitivity to misspecification in a fixed-T setting.

In panel data models, we show that minimizing the sensitivity to local misspecification of the distribution of individual effects leads to simple expressions, which coincide with known semi-parametric methods when the parameter of interest is point-identified and root-N consistently estimable (Bonhomme, 2012, Bonhomme and Davezies, 2017). In non-regular point-identified settings, minimizing mean squared error leads to a regularization approach, where the penalization reflects the amount of misspecification allowed for, which is itself calibrated based on a detection error probability. Our approach is still applicable when point identification fails, as we illustrate in a numerical exercise based on a dynamic probit model. Lastly, our method is not restricted to misspecification solely stemming from the distribution of individual effects, and we provide results when other aspects of the panel data model are 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 some neighborhoods 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 is that we focus on misspecification of specific aspects of a model, hence considering *semi-parametric* alternative models, while the robust statistics approach has mostly focused on data contamination and fully *nonparametric* alternatives.

A related branch of the literature is the work on locally robust moment functions, as developed in Neyman (1959), Newey (1994), and Chernozhukov *et al.* (2016), among others. Similarly to those approaches, we wish to construct estimators which are not too sensitive to

variation in an input. However, our analysis differs from the locally robust moment functions approach in that in our case the input is a (possibly non point-identified) model, as opposed to it being a parameter estimated in a first step. Another difference is that we take into account both bias and variance, weighting them by calibrating the size of the neighborhood of the reference model. 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), or recently Mueller (2012). In our approach we similarly focus on sensitivity to model, or "prior" assumptions. However, our goal is to minimize sensitivity and achieve minimum mean squared error in a frequentist sense.

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.

Lastly, this paper relates to the literature on sensitivity analysis in economics, for example Leamer (1985), Imbens (2003), Altonji *et al.* (2005), and recently Oster (2014). Our approach based on local misspecification has a number of precedents, such as Conley *et al.* (2012), Guggenberger (2012), Bugni *et al.* (2012), Kitamura *et al.* (2013), or recently Andrews *et al.* (2017). Chen *et al.* (2011) and Norets and Tang (2014) develop useful 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. We view our approach as complementary to these partial identification methods. Our local approach allows tractability in complex models, such as structural economic models. 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 studied by Hansen (2016) and Fessler and Kasy (2017).

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 present the main results. Extensions are shown in later sections (in progress).

2 Outline of the approach

We start by describing the main elements of our approach in a general setting. Let $f_\theta(y)$ be a model indexed by a finite or infinite-dimensional parameter θ. f_θ may be a discrete or continuous density, depending on the context. In most of the analysis we focus on a function, or functional, of θ, which we denote as δ_θ. The mapping $\theta \mapsto \delta_\theta$ is known, and we assume that δ_θ is scalar to keep the notation simple. The setup can easily be modified to allow for conditional models $f_\theta(y \,|\, x)$ given exogenous covariates, as we will do in our illustrations below, although here we abstract from x for conciseness. Examples of functionals of interest in economic applications include counterfactual policy effects which can be computed given a fully specified structural model, or moments of observed and latent data such as average effects in panel data settings.

We suppose that the researcher has a reference value for θ, denoted as θ^*. However, she only thinks of f_{θ^*} as a plausible approximation to the population distribution f_{θ_0}, and she is concerned about θ^* being misspecified. Our aim is to develop a framework to study the performance of estimators of δ_{θ_0} in case where the true θ_0 may differ from θ^*. Here we take θ^* to be a known, non-random parameter. We extend the framework to the case where θ^* is estimated in Section 7.

We study the properties of estimators of δ_{θ_0} of the form:

$$\widehat{\delta}_h = \frac{1}{N} \sum_{i=1}^{N} h(Y_i),$$

for a random sample $Y_1, ..., Y_N$ drawn from f_{θ_0}. The true value θ_0 is assumed to belong to a ball around θ^* with squared radius ϵ; that is, such that $\|\theta_0 - \theta^*\|^2 \leq \epsilon$. For reasons of tractability, especially in cases where θ includes infinite-dimensional components such as distribution functions, we focus on the case where the norm $\|\cdot\|$ is locally quadratic around zero; that is:

$$\|\theta_0 - \theta^*\|^2 \equiv \|\theta_0 - \theta^*\|_\Omega^2 = (\theta_0 - \theta^*)' \, \Omega \, (\theta_0 - \theta^*) + O\left(\|\theta_0 - \theta^*\|_\Omega^3\right),$$

where Ω is a non-singular weight matrix (or an operator when θ is infinite-dimensional). A leading example, which we will use in our three illustrations, is the Kullback-Leibler divergence between distribution functions.

We impose that, under correct specification $\theta_0 = \theta^*$, $\widehat{\delta}_h$ should be consistent for δ_{θ_0}. This

condition, which is sometimes referred to as "Fisher consistency", takes the following form:

$$\int h(y) f_{\theta^*}(y) dy = \delta_{\theta^*}. \tag{1}$$

In this setting, we define the minimax *bias*, or *sensitivity*, of $\widehat{\delta}_h$ as the largest distance between the true δ_{θ_0} and the probability limit of $\widehat{\delta}_h$, for θ_0 in an ϵ-neighborhood of θ^*; that is:

$$b_\epsilon(h, \theta^*) = \sup_{\theta_0 \in \Gamma_\epsilon(\theta^*)} \left| \delta_{\theta_0} - \int h(y) f_{\theta_0}(y) dy \right|, \tag{2}$$

where:

$$\Gamma_\epsilon(\theta^*) = \{\theta_0 \,:\, \|\theta_0 - \theta^*\|_\Omega^2 \leq \epsilon\}.$$

A *minimum sensitivity* estimator is based on a function h_θ^{MS} which minimizes $b_\epsilon(h, \theta^*)$ subject to (1). We will see that, under our local asymptotic framework, one may take $h_{\theta^*}^{MS}$ such that it does not depend on ϵ. To mitigate the variance increase associated with minimizing $b_\epsilon(h, \theta^*)$, we propose to also compute solutions to the following minimax *mean squared error* (MSE) problem:

$$\inf_h \left\{ \sup_{\theta_0 \in \Gamma_\epsilon(\theta^*)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_h - \delta_{\theta_0} \right)^2 \right], \quad s.t. \quad h \text{ satisfies (1)} \right\}. \tag{3}$$

We will denote solutions to (3), or a local asymptotic approximation to it, as $h_{\epsilon,\theta^*}^{MMSE}$, and will refer to $h_{\epsilon,\theta^*}^{MMSE}$ as a *minimum-MSE* estimator of δ_{θ_0}.

In general, solving (3) exactly for fixed ϵ and N is hard. This motivates us to take a *local asymptotic approach* and focus on an asymptotic where ϵ tends to zero as N tends to infinity. We will particularly focus on the case where the product $N\epsilon$ tends to a non-zero constant in the limit. This shrinking-neighborhood approach is a common device to reflect bias-variance trade-offs. Indeed, as ϵ tends to zero and N tends to infinity the MSE becomes, under standard regularity conditions:

$$\sup_{\theta_0 \in \Gamma_\epsilon(\theta^*)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_h - \delta_{\theta_0} \right)^2 \right] = \underbrace{b_\epsilon(h, \theta^*)^2}_{\text{squared bias}} + \underbrace{\frac{\text{Var}_{\theta^*}(h(Y))}{N}}_{\text{variance}} + o\left(N^{-1}\right) + o(\epsilon), \tag{4}$$

where the main simplification relative to the term in brackets in (3) comes from the fact that the variance is evaluated at the reference model θ^*. Under our local approach the squared bias b_ϵ^2 will be proportional to ϵ, so bias and standard deviation will be of the same asymptotic order as $N\epsilon$ tends to a constant.

In addition to point estimates, we will compute confidence intervals that contain δ_{θ_0} with prespecified probability. Based on any h function, for any confidence level μ we construct $CI_{\epsilon,\theta^*}(1-\mu,h)$ such that:

$$\liminf_{N\to\infty} \inf_{\theta_0\in\Gamma_\epsilon(\theta^*)} \Pr_{\theta_0}\left(\delta_{\theta_0} \in CI_{\epsilon,\theta^*}(1-\mu,h)\right) \geq 1-\mu.$$

To construct confidence intervals $CI_{\epsilon,\theta^*}(1-\mu,h)$, as well as to compute minimum-MSE moment functions $h^{MMSE}_{\epsilon,\theta^*}$, we need to set the neighborhood size ϵ. We follow a similar calibration approach as Hansen and Sargent (2008), who target the probability of a model detection error. Specifically, for a fixed probability p we choose $\epsilon(p)$ such that:

$$\liminf_{N\to\infty} \inf_{\theta_0\in\Gamma_{\epsilon(p)}(\theta^*)} \Pr_{\theta^*}\left(\sum_{i=1}^{N} \ln\left(\frac{f_{\theta_0}(Y_i)}{f_{\theta^*}(Y_i)}\right) > 0\right) \leq p. \tag{5}$$

Taking $\epsilon(p)$ such that (5) holds guarantees that, for some θ_0 in $\Gamma_{\epsilon(p)}(\theta^*)$, the probability of incorrectly detecting that θ_0 is more likely to have generated the data than θ^* is bounded by p. When p is fixed to a small value, say 1% or 5%, parameters outside the neighborhood $\Gamma_{\epsilon(p)}(\theta^*)$ will thus be "easy to detect" based on a sample of N observations. Achieving a lower p requires setting a higher $\epsilon(p)$. We will verify that, when p is kept fixed as N tends to infinity, and (5) holds with equality, then $N\epsilon(p)$ tends to a non-zero constant.

In summary, our framework delivers minimum-sensitivity estimators, based on $h^{MS}_{\theta^*}$, and collections of minimum-MSE estimators $h^{MMSE}_{\epsilon(p),\theta^*}$ and confidence intervals $CI_{\epsilon(p),\theta^*}(1-\mu,h)$ for different p levels. Reporting those allows one to assess sensitivity of the estimates to model misspecification and sampling uncertainty, and to construct improved estimates.

We now outline three examples that we will use as illustrations.

2.1 Example 1: Demand estimation

In the first example we consider a setting where there are J products. Individual i chooses product $Y_i = j$ if j maximizes her utility $U_{ij} = X_{ij}'\beta_j + \varepsilon_{ij}$, where X_{ij} are observed characteristics and ε_{ij} are random preference shocks; that is:

$$Y_i = j \Leftrightarrow X_{ij}'\beta_j + \varepsilon_{ij} \geq X_{ik}'\beta_k + \varepsilon_{ik} \text{ for all } k \neq j. \tag{6}$$

We assume that the vector of individual preference shocks $\varepsilon = (\varepsilon_1, ..., \varepsilon_J)$ is independent of $X = (X_1, ..., X_J)$, with density f_ε. We are interested in predictions from the demand

model, such has counterfactual market shares under different prices or other attributes of the goods. We denote such effects as:

$$\delta(\theta_0) = \mathbb{E}_{\theta_0}\left(\Delta(X, \varepsilon)\right),$$

for a known function Δ, where θ_0 denotes the true value of $\theta = f_\varepsilon$.

We start with a reference parametric specification $\theta^* = f^*_\varepsilon$. However we are concerned that the latter may be misspecified. This concern is widely echoed in the literature on demand analysis. A common example of a reference specification is ε_j being i.i.d. type-I extreme value, leading to a multinomial logit demand model. A common concern is that properties of the logit, in particular independence of irrelevant alternatives (IIA), may have unwanted consequences for the estimation of $\delta(\theta_0)$; see Anderson, De Palma and Thisse (1992), for example.

In this context, our approach allows us to perform a sensitivity analysis when the reference model is incorrect. We assess how the estimate of $\delta(\theta_0)$ changes under deviations from the reference model θ^* that matter most for our target parameter, and we modify the estimate in order to minimize the impact of such deviations.

2.2 Example 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, 1983). Our target parameters are $\delta^d = \mathbb{E}(Y(d))$, for $d = 0, 1$, and the average treatment effect $\delta^{(1)} - \delta^{(0)}$. This example is easily extended to the estimation of distribution of potential outcomes, beyond averages.

Let $\theta = f_d(y \,|\, x)$ denote the density of $Y(d)$ given $X = x$. We assume that the propensity score $p(x) = \Pr(D = 1 \,|\, x)$ is correctly specified. However, we allow for the reference θ^* to be misspecified. For example, our starting point may be a regression specification for $\mathbb{E}_*\left(Y(d) \,|\, X\right) = X'\beta^*_d$. While the estimator $\frac{1}{N}\sum_{i=1}^N X_i'\beta^*_d$ is consistent for $\delta^{(d)}$ under correct specification of the conditional mean, it becomes generally inconsistent when the specification $\mathbb{E}_*\left(Y(d) \,|\, X\right)$ is incorrect. Through our approach, we provide a sensitivity analysis to violations from that assumption, and show how to construct estimators of $\delta^{(d)}$, and the average treatment effect, which are less sensitive to misspecification.

2.3 Example 3: Average effects in panel data models

In our third example, which will be our main illustration in this paper, we consider panel data models with N cross-sectional units and T time periods. We allow for $T = 1$, in which case we effectively only have a cross-section of data, and more generally for any fixed $T \geq 1$. For each individuals $i = 1, \ldots, N$ 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. We assume that 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 α_i. Leaving i subscripts implicit for conciseness, we denote the corresponding probability density or probability mass function by $g(y \,|\, \alpha, x)$. In turn, the density of latent individual effects is denoted as $\pi(\alpha \,|\, x)$. The density of Y given X is then:

$$f_{g,\pi}(y \,|\, x) = \int g(y \,|\, \alpha, x)\pi(\alpha \,|\, x)d\alpha, \text{ for all } y, x.$$

The density of X, denoted as f_X, is left un-modeled. 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).

Here θ contains both g and π. We take as given a reference specification based on some g^* and π^*. In applications, g^* and π^* typically depend on parameters which are estimated. In Section 7 we will show how to take into account estimation uncertainty in these parameters within our approach. We will also study the question of the sensitivity and robustness to model misspecification of parameter estimates.

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

$$\begin{aligned}\delta(g_0, \pi_0) &= \mathbb{E}_{g_0,\pi_0,f_X}\left[\Delta(Y, A, X)\right] \\ &= \iiint \Delta(y, \alpha, x) g_0(y \,|\, \alpha, x)\pi_0(\alpha \,|\, x) f_X(x) dy d\alpha dx, \end{aligned} \tag{7}$$

for a known function Δ, where A denotes the random individual effect with (latent) realizations $\alpha_1, ..., \alpha_N$. 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).

Given g^*, π^*, an empirical counterpart to $\delta(g_0, \pi_0)$ is the *random-effects* estimator:

$$\widehat{\delta}^{RE} = \frac{1}{N}\sum_{i=1}^{N} \iint \Delta(y, \alpha, X_i) g^*(y \,|\, \alpha, X_i)\pi^*(\alpha \,|\, X_i) dy d\alpha. \tag{8}$$

Another commonly used estimator is *empirical Bayes* (or "Bayesian fixed-effects") estimator:

$$\widehat{\delta}^{EB} = \frac{1}{N}\sum_{i=1}^{N}\int \Delta(Y_i,\alpha,X_i)\underbrace{\frac{g^*(Y_i\,|\,\alpha,X_i)\pi^*(\alpha\,|\,X_i)}{\int g^*(Y_i\,|\,a,X_i)\pi^*(a\,|\,X_i)da}}_{=p^*(\alpha\,|\,Y_i,X_i)}d\alpha, \tag{9}$$

where $p^*(\alpha\,|\,Y_i,X_i)$ denotes the posterior distribution of individual effects α_i given (Y_i,X_i) implied by g^* and π^*. Note that $\widehat{\delta}^{EB}$ is the posterior mean of $\frac{1}{N}\sum_{i=1}^{N}\Delta(Y_i,\alpha_i,X_i)$ given the data, under the likelihood g^* and the prior π^*, both independent across individuals. Both $\widehat{\delta}^{RE}$ and $\widehat{\delta}^{EB}$ will be consistent for fixed T as N tends to infinity under correct specification. Our interest centers on situations where misspecification of π^* or g^* makes such commonly used estimators fixed-T inconsistent.

We will mostly focus on the case where g^* is correctly specified, and π^* may be misspecified. This case has received substantial attention in the literature, since misspecifying the distribution of unobserved heterogeneity and its dependence on initial conditions and other covariates can generate serious biases (Heckman, 1981). There is a large 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 $\widehat{\delta}^{RE}$, $\widehat{\delta}^{EB}$ generally remains consistent as both N and T tend to infinity when π^* is misspecified.

Finally, our local robustness approach allows us to consider other forms of model misspecification than the sole misspecification of the distribution of individual effects. We will provide additional results where either g^*, or both g^* and π^*, are misspecified.

3 Bias and minimum sensitivity

In this section, we characterize the form of the bias and minimum-sensitivity estimators in a local asymptotic framework where ϵ tends to zero as N tends to infinity, where the parameter θ may be finite or infinite-dimensional. We then apply our approach to our three examples, where θ is a distribution function, hence infinite-dimensional.

3.1 General case

Let $\nabla g_{\widetilde{\theta}}$ denotes the Gateaux derivative of g_θ with respect to θ evaluated at $\widetilde{\theta}$, and let $\|\cdot\|_{\Omega^{-1}}$ be the dual norm to $\|\cdot\|_\Omega$.[1] We have the following result, which we derive below. The proof

[1] That is, $\nabla g_{\widetilde{\theta}}(u) = \lim_{\tau\to 0}\frac{g_{\widetilde{\theta}+\tau u}-g_{\widetilde{\theta}}}{\tau}$, and $\|V\|^2_{\Omega^{-1}} = V'\Omega^{-1}V$.

requires suitable regularity conditions, which we will provide in the next version of the paper.

Proposition 1 *Let h be a moment function. The bias satisfies, as ϵ tends to zero and under suitable regularity conditions:*

$$b_\epsilon(h,\theta^*) = \epsilon^{\frac{1}{2}} \left\| \nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy \right\|_{\Omega^{-1}} + o\left(\epsilon^{\frac{1}{2}}\right). \tag{10}$$

Moreover, any function $h^{MS}_{\theta^*}$ satisfying the following, in addition to (1), is of minimum sensitivity:

$$\left[\nabla \ln f_{\theta^*}(y)\right]' \Omega^{-1} \left(\nabla\delta_{\theta^*} - \int h^{MS}_{\theta^*}(\widetilde{y})\nabla f_{\theta^*}(\widetilde{y})d\widetilde{y} \right) = 0, \text{ for all } y. \tag{11}$$

As an intuition for (11) note that, by (1) and expanding around θ^* as ϵ tends to zero, we have:

$$\delta_{\theta_0} = \mathbb{E}_{\theta_0}(h(Y)) + \left(\nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy \right)(\theta_0 - \theta^*) + o\left(\epsilon^{\frac{1}{2}}\right).$$

As a result, if $\nabla\delta_{\theta^*}$ and $\int h(y)\nabla f_{\theta^*}(y)dy$ are equal to each other, then the probability limit of $\widehat{\delta}_h$ is equal to the true value δ_{θ_0}, up to lower-order terms. Such a function h thus makes $\widehat{\delta}_h$ insensitive to local misspecification of θ^*. Such an estimator is thus "locally robust" according to the terminology of Chernozhukov *et al.* (2016). However, note that $h^{MS}_{\theta^*}$ in (11) need not exist, or be unique. Moreover, a solution to (11) need not have finite variance, especially in cases where θ is infinite-dimensional. These aspects of the minimum-sensitivity functions motivate studying mean squared errors minimizers, which are always unique and well-defined. We will study those in Section 4.

The formal justification for Proposition 1 can be described using variational arguments, which we outline here since most results we will describe in this paper will be based on a similar logic. The Lagrangian associated with the computation of the bias in (2) is:

$$\mathcal{L}_1 = \delta_{\theta_0} - \int h(y) f_{\theta_0}(y)dy + \lambda_1 \|\theta_0 - \theta^*\|^2_\Omega,$$

where λ_1 is a scalar Lagrange multiplier. Taking first-order conditions with respect to θ_0 gives:

$$\begin{aligned} 0 &= \nabla\delta_{\theta_0} - \int h(y)\nabla f_{\theta_0}(y)dy + 2\lambda_1\Omega(\theta_0 - \theta^*) \\ &= \nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy + 2\lambda_1\Omega(\theta_0 - \theta^*) + o\left(\epsilon^{\frac{1}{2}}\right). \end{aligned}$$

Hence:

$$\theta_0 - \theta^* = \pm\epsilon^{\frac{1}{2}} \frac{\Omega^{-1}\left(\nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy\right)}{\left\|\nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy\right\|_{\Omega^{-1}}} + o\left(\epsilon^{\frac{1}{2}}\right),$$

where we have used that the least-favorable θ_0 satisfies $\|\theta_0 - \theta^*\|_\Omega^2 = \epsilon$. Lastly:

$$\begin{aligned}\delta_{\theta_0} - \int h(y) f_{\theta_0}(y)dy &= \left(\nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy\right)(\theta_0 - \theta^*) + o\left(\epsilon^{\frac{1}{2}}\right) \\ &= \pm\epsilon^{\frac{1}{2}} \left\|\nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy\right\|_{\Omega^{-1}} + o\left(\epsilon^{\frac{1}{2}}\right),\end{aligned}$$

which implies (10).

Turning to (11), the Lagrangian associated with the minimization of $b_\epsilon(h, \theta^*)$ is:

$$\epsilon \left\|\nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy\right\|_{\Omega^{-1}}^2 + \lambda_2 \int h(y) f_{\theta^*}(y)dy,$$

where λ_2 is the Lagrange multiplier associated with (1). Taking first-order conditions with respect to h gives:

$$-2\epsilon\left[\nabla f_{\theta^*}(y)\right]'\Omega^{-1}\left(\nabla\delta_{\theta^*} - \int h(y)\nabla f_{\theta^*}(y)dy\right) + \lambda_2 f_{\theta^*}(y) = 0, \text{ for all } y.$$

Integrating with respect to y gives $\lambda_2 = 0$, and dividing by $f_{\theta^*}(y)$ gives (11).

3.2 Misspecification bias in Examples 1 and 2

Example 1 (continued). In the logit demand model the bias is:

$$b_\epsilon(h, \theta^*) = \epsilon^{\frac{1}{2}} \left\{\mathbb{E}_*\left[\left(\Delta(\varepsilon, X) - \sum_{j=1}^J q_j^*(\varepsilon, X)h(j, X)\right)^2\right]\right\}^{\frac{1}{2}} + o\left(\epsilon^{\frac{1}{2}}\right), \tag{12}$$

where:

$$q_j^*(\varepsilon, X) = \mathbf{1}\left\{X_{ij}'\beta_j + \varepsilon_{ij} \geq X_{ik}'\beta_k + \varepsilon_{ik} \text{ for all } k \neq j\right\}.$$

In (12), and throughout the paper, the $*$ subscript indicates an expectation taken with respect to the reference model f_{θ^*}. Any moment function $h_{\theta^*}^{MS}$ satisfying the following has minimum sensitivity:

$$\mathbb{E}_*\left[\Delta(\varepsilon, x) - \sum_{j=1}^J q_j^*(\varepsilon, x)h_{\theta^*}^{MS}(j, x) \;\middle|\; Y = k, x\right] = 0, \text{ for all } k = 1, ..., K \text{ and } x. \tag{13}$$

Example 2 (continued). Given functions $h^{(1)}$ and $h^{(0)}$, consider estimators of $\delta^{(1)}$ and $\delta^{(0)}$ given by $\widehat{\delta}_{h^{(1)}} = \frac{1}{N}\sum_{i=1}^{N} D_i h^{(1)}(Y_i, X_i)$, and $\widehat{\delta}_{h^{(0)}} = \frac{1}{N}\sum_{i=1}^{N}(1-D_i)h^{(0)}(Y_i, X_i)$. Their biases are, respectively:

$$b_\epsilon^{(1)}(h^{(1)}, \theta^*) = \epsilon^{\frac{1}{2}} \left\{ \mathbb{E}_* \left[\left(Y(1) - p(X) h^{(1)}(Y(1), X) \right)^2 \right] \right\}^{\frac{1}{2}} + o\left(\epsilon^{\frac{1}{2}}\right), \tag{14}$$

$$b_\epsilon^{(0)}(h^{(0)}, \theta^*) = \epsilon^{\frac{1}{2}} \left\{ \mathbb{E}_* \left[\left(Y(0) - [1-p(X)] h^{(0)}(Y(0), X) \right)^2 \right] \right\}^{\frac{1}{2}} + o\left(\epsilon^{\frac{1}{2}}\right). \tag{15}$$

The unique (up to lower-order terms) minimum-sensitivity moment functions are thus:

$$h_{\theta^*}^{MS,(1)}(y, x) = \frac{y}{p(x)}, \tag{16}$$

$$h_{\theta^*}^{MS,(0)}(y, x) = \frac{y}{1-p(x)}. \tag{17}$$

Note that estimators $\widehat{\delta}_{h^{(d)}}$ based on $h_{\theta^*}^{MS,(d)}$ are *inverse probability weighting* estimators, which remain consistent under misspecification of the distribution θ^* of potential outcomes, irrespective of the value of ϵ.

3.3 Misspecification bias in panel data

Given an i.i.d. sample $Y_1, X_1, ..., Y_N, X_N$ from (Y, X), we focus on an estimator of $\delta(g_0, \pi_0)$ in (7) of the form:

$$\widehat{\delta}_h = \frac{1}{N}\sum_{i=1}^{N} h(Y_i, X_i).$$

We impose that the estimator is consistent under correct specification for any covariates distribution; that is:

$$\iint \left(\Delta(y, \alpha, x) - h(y, x)\right) g^*(y \,|\, \alpha, x)\pi^*(\alpha \,|\, x) dy d\alpha = 0, \text{ for all } x. \tag{18}$$

For given $\epsilon > 0$, we consider the following set $\Gamma_\epsilon(g^*, \pi^*)$ for (g_0, π_0):

$$\begin{aligned} &\int \pi_0(\alpha \,|\, x) d\alpha = 1 \text{ for all } x, \\ &\int g_0(y \,|\, \alpha, x) dy = 1 \text{ for all } (\alpha, x), \\ &d_{KL}(g_0\pi_0 f_X, g^*\pi^* f_X) \leq \epsilon, \end{aligned}$$

where $\frac{1}{2}d_{KL}$ denotes the Kullback-Leibler divergence.[2] Notice that $d_{KL}(g_0\pi_0 f_X, g^*\pi^* f_X)$ is expressed in terms of *joint* distributions.

[2]Formally, for two densities f_0 and f of (y, α, x), $d_{KL}(f_0, f) = 2\iiint \ln\left(\frac{f_0(y,\alpha,x)}{f(y,\alpha,x)}\right) f_0(y, \alpha, x) dy d\alpha dx$. Adding the factor 2 simplifies some of the expressions below.

The bias of $\widehat{\delta}_h$ is then :

$$b_\epsilon(g^*, \pi^*, h) = \sup_{(g_0,\pi_0)\in\Gamma_\epsilon(g^*,\pi^*)} \left| \delta(g_0, \pi_0) - \iint h(y, x) f_{g_0,\pi_0}(y \,|\, x) f_X(x) dy dx \right|.$$

3.3.1 Misspecified π^*

We start by focusing on the case where $g^* = g_0$ is correctly specified, misspecification only coming from $\pi^* \neq \pi^0$. As a corollary to Proposition 1, the next result establishes the form of the bias, and gives the expression for minimum-sensitivity estimators of $\delta(g_0, \pi_0)$, in a local asymptotic where ϵ tends to zero. Here again, the $*$ subscript means that expectations and variances are taken with respect to the density of the reference model indexed by g^*, π^*.

Corollary 1 *Consider the panel data setup under the assumption that $g^0 = g^*$ is correctly specified. Then, as ϵ tends to zero:*

$$b_\epsilon(g^*, \pi^*, h) = \epsilon^{\frac{1}{2}} \left\{ \mathrm{Var}_* \left[\mathbb{E}_* \left(\Delta(Y, A, X) - h(Y, X) \,|\, A, X \right) \right] \right\}^{\frac{1}{2}} + o\left(\epsilon^{\frac{1}{2}}\right).$$

Moreover, any function $h^{MS}_{g^*,\pi^*}$ satisfying the following is of minimum sensitivity:

$$\mathbb{E}_* \left[\mathbb{E}_* \left(\Delta(Y, A, x) - h^{MS}_{g^*,\pi^*}(Y, x) \,|\, A, x \right) \,|\, y, x \right] = 0, \text{ for all } y, x.$$

The bias in Corollary 1 is useful to assess the sensitivity of a given estimator to misspecification of π^*. As an example, consider the random-effects and empirical Bayes estimators $\widehat{\delta}^{RE}$ and $\widehat{\delta}^{EB}$ given by (8) and (9). Abstracting from lower-order terms, the bias for the random-effects estimator is:

$$b^{RE}_\epsilon(g^*, \pi^*, h) = \epsilon^{\frac{1}{2}} \left\{ \mathrm{Var}_* \left[\Delta(Y, A, X) \right] \right\}^{\frac{1}{2}},$$

while the bias for the empirical Bayes estimator is:

$$b^{EB}_\epsilon(g^*, \pi^*, h) = \epsilon^{\frac{1}{2}} \left\{ \mathrm{Var}_* \left[\mathbb{E}_* \left(\Delta(Y, A, X) - \mathbb{E}_* \left[\Delta(Y, \widetilde{A}, X) \,|\, Y, X \right] \,\middle|\, A, X \right) \right] \right\}^{\frac{1}{2}}.$$

When π^* is misspecified, neither random-effects nor empirical Bayes estimators minimize misspecification bias in general. However, the bias result gives a clear ranking between the two estimators in terms of sensitivity. It is easy to see that $b^{RE} \geq b^{EB}$, with strict inequality unless the conditional mean of $\Delta(Y, A, X)$ given (Y, X) is constant. In addition, in the case

where $\Delta(Y, A, X) = \Delta(A, X)$ is solely a function of individual effects and covariates, b^{RE} does not vary with T. In contrast, we expect that, as T tends to infinity, b^{EB} tends to zero.[3] 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).

Sensitivity, identification, and fixed-T consistency. To get insight on the form of the minimum-sensitivity estimator in Corollary 1, let us focus for simplicity on the case where $\Delta(Y, A, X) = \Delta(A, X)$ does not depend on Y. Consider the case where there exists a function h_Δ such that:

$$\mathbb{E}_* \left[h_\Delta(Y, x) \,|\, \alpha, x \right] = \Delta(\alpha, x), \quad \text{for all } \alpha, x.$$

The existence of such an h_Δ, under suitable regularity conditions, is necessary and sufficient for $\delta(g_0, \pi_0)$ to be identified and root-N consistently estimable for fixed T (Bonhomme and Davezies, 2017). In this case, by Corollary 1, $h^{MS}_{g^*,\pi^*} - h_\Delta$ satisfies:

$$\mathbb{E}_* \left[h^{MS}_{g^*,\pi^*}(Y, x) - h_\Delta(Y, x) \,|\, \alpha, x \right] = 0, \quad \text{for all } \alpha, x.$$

We thus have:

$$\begin{aligned}
\mathbb{E}_{g_0,\pi_0,f_X} \left[h^{MS}_{g^*,\pi^*}(Y, X) \right] &= \mathbb{E}_{g_0,\pi_0,f_X} \left[h_\Delta(Y, X) \right] + \mathbb{E}_{g_0,\pi_0,f_X} \left[h^{MS}_{g^*,\pi^*}(Y, X) - h_\Delta(Y, X) \right] \\
&= \mathbb{E}_{\pi_0,f_X} \left[\Delta(A, X) \right] + \mathbb{E}_{\pi_0,f_X} \left[\underbrace{\mathbb{E}_* \left[h^{MS}_{g^*,\pi^*}(Y, X) - h_\Delta(Y, X) \,|\, A, X \right]}_{=0} \right] \\
&= \mathbb{E}_{\pi_0,f_X} \left[\Delta(A, X) \right].
\end{aligned}$$

As a result, the sample mean of $h^{MS}_{g^*,\pi^*}(y_i, x_i)$ is a fixed-T, root-N consistent estimator of $\delta(g_0, \pi_0)$ *irrespective* of the value of ϵ, provided $h^{MS}_{g^*,\pi^*}(Y, X)$ has finite first moment. In this case, ensuring local in-sensitivity leads to full robustness against misspecification of π^*.

Existence of a function h_Δ is a strong condition. In particular, it presumes that $\delta(g_0, \pi_0)$ 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,

[3] Indeed we have, as T tends to infinity:

$$\mathbb{E} \left[\mathbb{E} \left(\Delta(A, x) \,|\, Y, x \right) \,|\, \alpha, x \right] \approx \mathbb{E} \left(\widehat{\alpha}(Y, x) \,|\, \alpha, x \right) \approx \Delta(\alpha, x), \text{ for all } \alpha, x,$$

where $\widehat{\alpha}(y, x) = \text{argmax}_a \, g^*(y \,|\, \alpha, x)$ is the maximum likelihood estimator of α (for a given individual).

2006, Chernozhukov *et al.*, 2013, Pakes and Porter, 2013). When $\delta(g_0, \pi_0)$ is point-identified, there will exist a sequence $h_{\Delta,N}$ such that the conditional mean $\mathbb{E}_* [h_{\Delta,N}(Y, x) \,|\, \alpha, x]$ converges to $\Delta(\alpha, x)$ as N tends to infinity in a suitable topology. In this case, a regularized version of the sample mean of $h^{MS}_{g^*,\pi^*}(y_i, x_i)$ will be consistent for $\delta(g_0, \pi_0)$, although it will not be root-N consistent. The minimum-MSE moment function $h^{MMSE}_{\epsilon,g^*,\pi^*}$, which we derive in the next section, can be interpreted as a regularized version of $h^{MS}_{g^*,\pi^*}$, with $N\epsilon$ acting as a regularization parameter.

Global versus local approaches. The expressions of the bias and minimum-sensitivity moment function in Corollary 1 are derived under ϵ tending to zero. The next result characterizes the bias for fixed ϵ.

Corollary 2 *Consider the panel data setup under the assumption that $g^0 = g^*$ is correctly specified. Then, for fixed ϵ:*

$$b_\epsilon(g^*, \pi^*, h) = \epsilon^{\frac{1}{2}} \Bigg\{ C \cdot \mathbb{E}_* \Bigg[\mathbb{E}_* \left(\Delta(Y, A, X) - h(Y, X) \,|\, A, X \right) \times \exp\left(-\frac{1}{2\lambda_2} \cdot \mathbb{E}_* \left(\Delta(Y, A, X) - h(Y, X) \,|\, A, X \right) \right) \Bigg] \Bigg\}^{\frac{1}{2}},$$

for two constants $C > 0$ and λ_2 which satisfy (A2)-(A3) in the appendix.

Corollary 2 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 it can be verified that $1/\lambda_2$ tends to zero, so the bias becomes proportional (in fact, equal) to the expression appearing in Corollary 1.

While it would be theoretically possible to follow a global approach throughout the analysis, instead of the local approach we advocate, proceeding in that way would face three main challenges. First, the bias in Corollary 2 depends on parameters C and λ_2 which need to be recovered given ϵ, increasing computational cost. Second, derivations such as the one in Corollary 2 are limited to settings where the parameter θ_0 (that is, π_0 in the panel setting we focus on here) enters the likelihood function linearly. Under linearity, similar derivations have been proven useful in other contexts; see for example Schennach (2013) for a recent example. The third and main challenge for implementing a global approach is that characterizing mean squared errors and confidence intervals would become less tractable, while as we will see below 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 the discussion following Corollary 1.

3.3.2 Misspecified g^*

We now consider panel data settings where $g^*(y \,|\, \alpha, x)$ may be misspecified. The following result characterizes the bias an minimum-sensitivity estimator when neither g^* nor π^* are assumed to be correctly specified.

Corollary 3 *Consider the panel data setup when neither g^* nor π^* are correctly specified. We have, as ϵ tends to zero:*

$$b_\epsilon(g^*, \pi^*, h) = \epsilon^{\frac{1}{2}} \left\{ \mathrm{Var}_* \Big[\Delta(Y, A, X) - h(Y, X) \Big] \right\}^{\frac{1}{2}} + o\left(\epsilon^{\frac{1}{2}}\right).$$

Moreover, the minimum-sensitivity moment function is unique (up to lower-order terms), and corresponds to the empirical Bayes estimator of $\delta(g_0, \pi_0)$; that is:

$$h^{MS}_{g^*,\pi^*}(y, x) = \mathbb{E}_* \big[\Delta(y, A, x) \,|\, y, x \big], \text{ for all } y, x.$$

Corollary 3 provides a characterization of the empirical Bayes estimator as the unique minimum-sensitivity estimator under any possible misspecification of the panel data model. Note that, in this case, there is no scope for achieving fixed-T or even large-T identification (except in the trivial case where $\Delta(Y, A, X) = \Delta(Y, X)$ does not depend on A).

The last result in this section gives the bias and minimum-sensitivity moment function when π^* is correctly specified, but g^* may be misspecified.

Corollary 4 *Consider the panel data setup when $\pi_0 = \pi^*$ is correctly specified. We have, as ϵ tends to zero:*

$$b_\epsilon(g^*, \pi^*, h) = \epsilon^{\frac{1}{2}} \left\{ \mathrm{Var}_* \left[\Delta(Y, A, X) - h(Y, X) - \mathbb{E}\left[\Delta(\widetilde{Y}, A, X) - h(\widetilde{Y}, X) \,|\, A, X \right] \right] \right\}^{\frac{1}{2}} + o\left(\epsilon^{\frac{1}{2}}\right).$$

Moreover, any function $h^{MS}_{g^*,\pi^*}$ satisfying the following minimizes sensitivity:

$$\mathbb{E}_* \left[\Delta(y, A, x) - h^{MS}_{g^*,\pi^*}(y, x) - \mathbb{E}\left(\Delta(Y, A, x) - h^{MS}_{g^*,\pi^*}(Y, x) \,|\, A, x \right) \,|\, y, x \right] = o(1), \text{ for all } y, x.$$

As an example, if $\Delta(Y, A, X) = \Delta(Y, X)$ does not depend on individual effects, then $h^{MS}_{g^*,\pi^*}(y, x) \equiv \Delta(y, x)$ minimizes the MSE, and the sample mean of $\Delta(y_i, x_i)$ is (obviously) fully robust. More generally, while in Corollary 1 the bias and minimum-sensitivity functions depend on between-α variances and expectations, in Corollary 4 they depend on within-α quantities.

4 Mean squared error and confidence intervals

In this section we provide characterizations for minimum-MSE estimators, confidence intervals which account for model misspecification, and our rule to calibrate the size $\epsilon = \epsilon(p)$ of the neighborhood around the reference model. We start with the general framework, and then provide results for our three illustrations.

4.1 Minimum-MSE estimation: the general case

We have the following result.

Proposition 2 *Minimizing the sum of squared-bias and variance on the right-hand side of (4), with respect to h subject to (1), we obtain, as N tends to infinity and ϵ tends to zero such that $N\epsilon$ tends to a non-zero constant:*

$$h(y) = \delta_{\theta^*} + N\epsilon \left[\nabla \ln f_{\theta^*}(y)\right]' \Omega^{-1} \left(\nabla \delta_{\theta^*} - \int h(y) \nabla f_{\theta^*}(y) dy\right) + o(1), \text{ for all } y.$$

We define the minimum-MSE moment function as solution to:

$$h^{MMSE}_{\epsilon,\theta^*}(y) = \delta_{\theta^*} + N\epsilon \left[\nabla \ln f_{\theta^*}(y)\right]' \Omega^{-1} \left(\nabla \delta_{\theta^*} - \int h^{MMSE}_{\epsilon,\theta^*}(y) \nabla f_{\theta^*}(y) dy\right), \text{ for all } y. \tag{19}$$

Proposition 2 shows that the minimum-MSE moment function solves a linear integral equation. An explicit expression for the solution is given by:

$$h^{MMSE}_{\epsilon,\theta^*}(y) = \delta_{\theta^*} + \left[\nabla \ln f_{\theta^*}(y)\right]' \left[H_{\theta^*} + (N\epsilon)^{-1}\Omega\right]^{-1} \nabla \delta_{\theta^*}, \tag{20}$$

where $H_{\theta^*} = \mathbb{E}_{\theta^*}\left[\nabla \ln f_{\theta^*}(Y) \nabla \ln f_{\theta^*}(Y)'\right]$, and $^{-1}$ denotes the matrix (or operator) inverse.

The presence of the $(N\epsilon)^{-1}$ term in (19) acts as a regularization device. It may be that H_{θ^*} is singular, or H_θ is singular for some θ in $\Gamma_\epsilon(\theta^*)$. This situation may arise when θ_0 is

not point-identified, for example. A regular inverse of H_{θ^*} may then fail to exist, or become ill-conditioned as ϵ tends to zero. The addition of the term $(N\epsilon)^{-1}\Omega$ makes the inverse well-defined. Moreover, even under point-identification, in cases where θ is infinite-dimensional such as in panel data models, the regularization helps to alleviate ill-posedness issues.

To see why Proposition 2 holds, it is useful to rely on a variational argument, as in Proposition 1. The Lagrangian associated with the constrained minimization of the variance in (4) is:

$$\mathcal{L} = \epsilon \left\| \nabla \delta_{\theta^*} - \int h(y) \nabla f_{\theta^*}(y) dy \right\|^2_{\Omega^{-1}} + \frac{1}{N} \int h^2(y) f_{\theta^*}(y) dy + \lambda \int h(y) f_{\theta^*}(y) dy,$$

where λ denotes the Lagrange multiplier associated with (1), and we have used the expression for the bias from Proposition 1. Taking first-order conditions with respect to h gives:

$$-2\epsilon \left[\nabla f_{\theta^*}(y)\right]' \Omega^{-1} \left(\nabla \delta_{\theta^*} - \int h(y) \nabla f_{\theta^*}(y) dy \right) + \frac{2}{N} h(y) f_{\theta^*}(y) + \lambda f_{\theta^*}(y) = 0, \text{ for all } y.$$

Integrating with respect to y, and using (1), the Lagrange multiplier is determined as:

$$\lambda = -\frac{2}{N} \delta_{\theta^*}.$$

Lastly, dividing by $f_{\theta^*}(y)$ implies Proposition 2.

4.2 Confidence intervals and calibration of ϵ

Let h be a moment function, and let: $\delta_h(\theta) = \int h(y) f_\theta(y) dy$ be the probability limit of $\widehat{\delta}_h$ under f_θ. Let c_μ denote the μ-th quantile of the standard normal. We have, for all confidence level μ in the unit interval and with probability approaching one:

$$\inf_\theta \Pr_\theta \left[\delta_h(\theta) - \frac{\sigma_h(\theta)}{\sqrt{N}} c_{1-\mu/2} \leq \widehat{\delta}_h \leq \delta_h(\theta) + \frac{\sigma_h(\theta)}{\sqrt{N}} c_{1-\mu/2} \right] \geq 1 - \mu,$$

where $\sigma^2_h(\theta) = \text{Var}_\theta\left(h(Y)\right)$.

It thus follows that, for given θ^* and with probability approaching one:

$$\inf_{\theta_0 \in \Gamma_\epsilon(\theta^*)} \Pr_{\theta_0} \left[\widehat{\delta}_h - b_\epsilon(h, \theta^*) - \frac{\sigma_h(\theta_0)}{\sqrt{N}} c_{1-\mu/2} \leq \delta(\theta_0) \leq \widehat{\delta}_h + b_\epsilon(h, \theta^*) + \frac{\sigma_h(\theta_0)}{\sqrt{N}} c_{1-\mu/2} \right] \geq 1 - \mu.$$

Hence, an asymptotically valid confidence interval for $\delta(\theta_0)$, asymptotically uniform on $\Gamma_\epsilon(\theta^*)$, can be constructed as:

$$\left[\widehat{\delta}_h \pm \left(\epsilon^{\frac{1}{2}} \left\| \nabla \delta_{\theta^*} - \int h(y) \nabla f_{\theta^*}(y) dy \right\| + \frac{\widehat{\sigma}_h}{\sqrt{N}} c_{1-\mu/2} \right) \right], \tag{21}$$

where $\widehat{\sigma}^2_h = \widehat{\text{Var}}(h(Y_i))$ is the sample variance of $h(Y_i)$. This shows that accounting for misspecification of θ^* leads to an enlargement of the classical $(1-\mu)$ confidence interval $\left[\widehat{\delta}_h \pm \frac{\widehat{\sigma}_h}{\sqrt{N}} c_{1-\mu/2} \right]$.

Calibration of ϵ. In order to obtain a meaningful bias-variance trade-off we focus on an asymptotic where $N\epsilon$ tends to a constant. To calibrate that constant we follow Hansen and Sargent (2008) and use an approach based on calibrating a detection error probability. Specifically, for a fixed probability p we choose $\epsilon(p)$ such that (5) holds.

We have, with probability approaching one as N and ϵ^{-1} tend to infinity such that $N\epsilon$ tends to a constant $\rho > 0$:

$$\inf_{\theta_0 \in \Gamma_\epsilon(\theta^*)} \Pr_{\theta^*} \left(\sum_{i=1}^{N} \ln \left(\frac{f_{\theta_0}(Y_i)}{f_{\theta^*}(Y_i)} \right) > 0 \right) = \Phi \left(-\frac{1}{2} \sqrt{\rho \cdot \lambda_{max}} \right),$$

where Φ is the standard Gaussian cdf, and λ_{max} is the maximal eigenvalue of the Hessian matrix (or operator in infinite-dimensional cases) $H_{\theta^*} = \mathbb{E}_{\theta^*} [\nabla \ln f_{\theta^*}(Y) \nabla \ln f_{\theta^*}(Y)']$.

This motivates us to take:

$$\epsilon(p) = \frac{4 \left(\Phi^{-1}(p) \right)^2}{N \cdot \lambda_{max}}. \tag{22}$$

4.3 Minimum MSE in Examples 1 and 2

Example 1 (continued). In the logit demand example, any moment function satisfying the following is of minimum MSE:

$$\begin{aligned} h^{MMSE}_{\epsilon,\theta^*}(k,x) &= \mathbb{E}_* \left[\Delta(\varepsilon, x) \;\middle|\; X = x \right] \\ &+ N\epsilon \, \mathbb{E}_* \left[\Delta(\varepsilon, x) - \sum_{j=1}^{J} q^*_j(\varepsilon, x) h^{MMSE}_{\epsilon,\theta^*}(j,x) \;\middle|\; Y = k, X = x \right], \text{ for all } k = 1, ..., K \text{ and } x. \end{aligned} \tag{23}$$

The function $h^{MMSE}_{\epsilon,\theta^*}$ can be approximated numerically by drawing from the posterior distribution of ε given (Y, X), for example using Markov Chain Monte Carlo techniques.

Example 2 (continued). The unique (up to lower-order terms) minimum-MSE moment functions are:

$$h^{MMSE,(1)}_{\epsilon,\theta^*}(y,x) = \frac{1}{1 + N\epsilon p(x)} \frac{\mathbb{E}_*[Y(1) \,|\, X = x]}{p(x)} + \frac{N\epsilon p(x)}{1 + N\epsilon p(x)} \frac{y}{p(x)}, \tag{24}$$

$$h^{MMSE,(0)}_{\epsilon,\theta^*}(y,x) = \frac{1}{1 + N\epsilon(1 - p(x))} \frac{\mathbb{E}_*[Y(0) \,|\, X = x]}{1 - p(x)} + \frac{N\epsilon(1 - p(x))}{1 + N\epsilon(1 - p(x))} \frac{y}{1 - p(x)}. \tag{25}$$

4.4 Minimum MSE in panel data

Here we provide results on minimum-MSE estimation for the panel data models of Example 3, in the case where g^* is correctly specified. Analogous results can easily be derived in case where either g^*, or both g^* and π^*, may be misspecified.

Corollary 5 *Consider the panel data setup under the assumption that $g_0 = g^*$ is correctly specified. Then, as N tends to infinity and ϵ tends to zero such that $N\epsilon$ tends to a non-zero constant, any moment function $h^{MMSE}_{\epsilon,g^*,\pi^*}$ solving the following linear system is of minimum MSE:*

$$h^{MMSE}_{\epsilon,g^*,\pi^*}(y,x) = \mathbb{E}_*\left[\Delta(Y,A,x)\,|\,x\right] + N\epsilon\,\mathbb{E}_*\left[\mathbb{E}_*\left(\Delta(\widetilde{Y},A,x) - h^{MMSE}_{\epsilon,g^*,\pi^*}(\widetilde{Y},x)\,|\,A,x\right)\,|\,y,x\right],\ \textit{for all } (y,x).$$

To provide intuition about Corollary 5, let us introduce the following conditional expectations operators, mapping squared-integrable functions:[4]

$$Lm(y,x) = \mathbb{E}_*\left(m(A,x)\,|\,y,x\right),$$
$$L^\star h(\alpha,x) = \mathbb{E}_*\left(h(Y,x)\,|\,\alpha,x\right).$$

From Corollary 5, $h^{MMSE}_{\epsilon,g^*,\pi^*}$ solves:

$$\left(I + N\epsilon LL^\star\right) h^{MMSE}_{\epsilon,g^*,\pi^*} = \mathbb{E}_*\left[\Delta(Y,A,x)\,|\,x\right] + N\epsilon\,\mathbb{E}_*\left(\Delta(y,A,x)\,|\,y,x\right),$$

where I is the identity mapping. As a result:

$$h^{MMSE}_{\epsilon,g^*,\pi^*}(y,x) = \frac{1}{1+N\epsilon}\mathbb{E}_*\left[\Delta(Y,A,x)\,|\,x\right] + \frac{N\epsilon}{1+N\epsilon}\left(\frac{I+N\epsilon LL^\star}{1+N\epsilon}\right)^{-1}\mathbb{E}_*\left(\Delta(y,A,x)\,|\,y,x\right). \tag{26}$$

Equation (26) shows that the minimum-MSE estimator of $\delta(g_0,\pi_0)$ is a weighted average of the random-effects estimator $\widehat{\delta}^{RE}$ and a regularized nonparametric estimator of $\delta(g_0,\pi_0)$ based on:

$$h^{NP}(y,x) \equiv \left(\frac{I+N\epsilon LL^\star}{1+N\epsilon}\right)^{-1}\mathbb{E}_*\left(\Delta(y,A,x)\,|\,y,x\right).$$

When $\delta(g_0,\pi_0)$ is point-identified, the mean of $h^{NP}(Y_i)$ will converge in probability to $\delta(g_0,\pi_0)$ as N tends to infinity (and for fixed T), provided $N\epsilon$ tends to infinity.[5] Due

[4] The notation $L^\star$ refers to the fact that L and $L^\star$ are adjoint operators with respect to the L^2 topology.

[5] In that case:

$$h^{NP} = \left(\frac{I+N\epsilon LL^\star}{1+N\epsilon}\right)^{-1} L\Delta$$

to ill-posedness, the behavior of the sample mean of $h^{NP}(Y_i)$ will typically be unstable, however, and depend heavily on the choice of $N\epsilon$ (e.g., Carrasco, Florens and Renault, 2007, Bonhomme, 2012). In contrast, calibrating ϵ as we do, based on a fixed error detection probability p, implies that the worst-case bias of the minimum-MSE estimator is controlled while keeping the problem well-posed. In addition, our approach still delivers a well-defined quantity in cases where the population parameter of interest is partially identified.

Note also that confidence intervals constructed as in (21), for the expression of the bias appearing in Corollary 1, will cover the true value $\delta(g_0, \pi_0)$ with pre-specified probability $(1 - \mu)$.

Calibration of ϵ. Finally, in the panel data case (22) becomes:

$$\epsilon = \frac{4\left(\Phi^{-1}(\mu)\right)^2}{N}, \tag{27}$$

where we have used that here λ_{max} is the maximum singular value of the operator:

$$LL^{\star}h(y,x) = \mathbb{E}_*\left[\left(\mathbb{E}\left[h(Y,x)\,|\,A,x\right]\right)\,|\,y,x\right], \text{ for all } y,x,$$

which is a doubly stochastic operator.

5 Estimating θ^* using a parametric submodel

We have assumed so far that θ^* is known, which is very useful for expositional simplicity, but is not a realistic assumption in most applications. This section discusses the case where θ^* is estimated using a low-dimensional sub-model. Let $\theta = \theta(\eta)$ be a known function of a finite dimensional parameter vector η, and let

$$\widehat{\eta} = \underset{\eta}{\operatorname{argmax}} \sum_{i=1}^{N} \log f_{\theta(\eta)}(Y_i|X_i), \qquad \eta^*(\theta_0) = \underset{\eta}{\operatorname{argmax}}\, \mathbb{E}_{\theta_0} \log f_{\theta(\eta)}(Y|X),$$

be the corresponding maximum likelihood estimator $\widehat{\eta}$ and pseudo-true parameter $\eta^*(\theta_0)$. If θ_0 is the true parameter, then we have $\widehat{\eta} \to_p \eta^*(\theta_0)$, as $n \to \infty$, under standard regularity conditions. We essentially want to replace θ_* by $\theta(\widehat{\eta})$ in our discussion of the minimum-MSE estimation problem, but in doing so we need to account for the sampling noise in $\theta(\widehat{\eta})$ as well, and we need to find a logically consistent formulation of the MSE minimization problem that allows for the limit of $\widehat{\eta}$ to depend on θ_0.

converges to $L^{\dagger}\Delta$ as $(N\epsilon)^{-1}$ tends to zero, where $L^{\dagger}$ denotes the Moore-Penrose generalized inverse of the conditional expectation operator L.

6 Numerical calculations for a dynamic probit model

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

$$Y_{it} = \mathbf{1}\left\{\beta_0 Y_{i,t-1} + \alpha_i + U_{it} \geq 0\right\}, \quad t = 2, ..., T,$$

where $U_{i2}, ..., U_{iT}$ are i.i.d. standard Gaussian, independent of Y_{i1} and α_i. We assume that the probit conditional likelihood given individual effects and lagged outcomes is correctly specified, and that β_0 is known. We take π^*, the density of α_i given Y_{i1}, to be Gaussian with mean $a_0 + b_0 Y_{i1}$ and variance σ_0^2, where a_0, b_0, σ_0 are fixed and known.

We focus on the average *state dependence* effect:

$$\delta_0 = \mathbb{E}_{\pi_0}\left[\Phi(\beta_0 + \alpha_i) - \Phi(\alpha_i)\right],$$

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

$$\widehat{\delta}^{RE} = \frac{1}{N}\sum_{i=1}^{N} \int \left[\Phi(\beta_0 + \alpha_i) - \Phi(\alpha_i)\right] \pi^*(\alpha_i \,|\, Y_{i1}) d\alpha_i.$$

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

$$\widehat{\delta}^{EB} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_*\left[\Phi(\beta_0 + \alpha_i) - \Phi(\alpha_i) \,|\, Y_{i1}, ..., Y_{iT}\right].$$

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

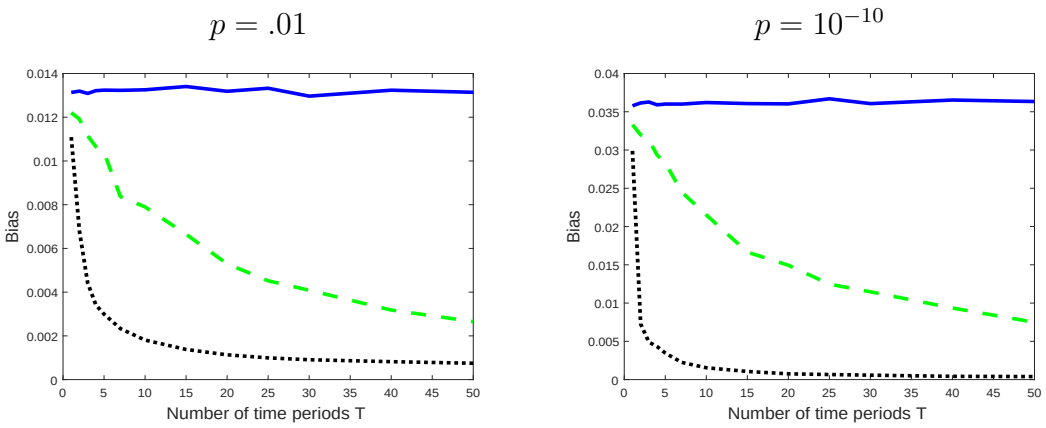
$$\widehat{\delta}_\epsilon^{MMSE} = \frac{1}{N}\sum_{i=1}^{N} h_{\epsilon,\pi^*}^{MMSE}(Y_{i1}, ..., Y_{iT}),$$

where $h_{\epsilon,\pi^*}^{MMSE}$ solves:

$$h_{\epsilon,\pi^*}^{MMSE}(y_1, ..., y_T) = \int \left[\Phi(\beta_0 + \alpha) - \Phi(\alpha)\right] \pi^*(\alpha \,|\, y_1) d\alpha$$
$$+ N\epsilon\, \mathbb{E}_*\Big[\Phi(\beta_0 + A) - \Phi(A) - \mathbb{E}_*\left(h_{\epsilon,\pi^*}^{MMSE}(y_1, Y_2, ..., Y_T) \,|\, A, y_1\right) \,|\, y_1, ..., y_T\Big], \text{ for all } y_1, ..., y_T.$$

We fix values for $\beta_0 = 1$, $a_0 = -.2$, $b_0 = 0$, and $\sigma_0 = .8$. In the simulation Y_{i1} is fixed to 0. We compute h^{MMSE} by sampling from π^* and g^*, 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 of different estimators of the average state dependence effect in the dynamic probit model

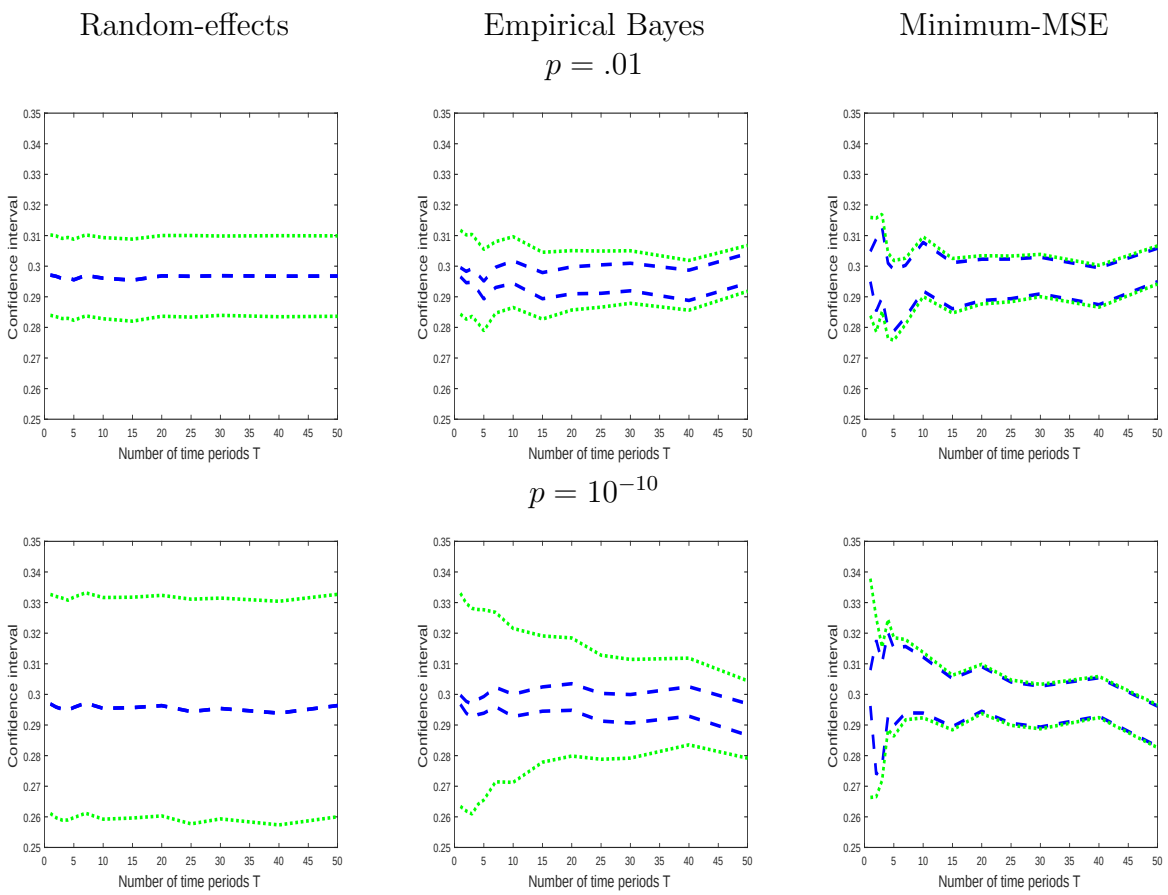


Notes: Bias b_ϵ for different panel length T. The solid line corresponds to the random-effects estimator $\widehat{\delta}^{RE}$, the dashed line to the empirical Bayes estimator $\widehat{\delta}^{EB}$, and the dotted line to the minimum-MSE estimator $\widehat{\delta}^{MMSE}$. ϵ is chosen according to (27) for a detection error probability $p = .01$ (left) and $p = 10^{-10}$ (right) when $N = 1000$.

In Figure 1 we show the bias b_ϵ for each of the three estimators, where ϵ is set according to (27) for a detection error probability $p = .01$ (left graph) and $p = 10^{-10}$ (right). 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 minimum-MSE estimator (dotted) has the smallest bias, and it decreases quickly as T increases. The bias levels off in the large-T limit, since the size of ϵ is indexed by N and independent of T. Lastly, setting p to a (much) smaller value implies larger biases for the random-effects and empirical Bayes estimators.

In Figure 2 we report two types of 95% confidence intervals for the average state dependence effect: obtained under correct specification (dashed lines), and allowing for local misspecification as in (21) (dotted lines). The number of individuals is $N = 1000$, and ϵ is chosen based on (27) for a rejection probability $p = .01$. We see that the concern for misspecification leads to enlarging the 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.

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



Notes: 95%-confidence intervals for the average state dependence effect, based on three estimators. Dashed lines correspond to confidence intervals based on correct specification, dotted lines to the ones allowing for local misspecification. $N = 1000$. ϵ *is chosen according to (27) for a detection error probability* $p = .01$ *(top) and* $p = 10^{-10}$ *(bottom).*

7 Extension: parametric approximating models

To be completed.

8 Extension: local equivalence to worst-case bounds

Consider the following *restricted identified set* for δ_{θ_0}, where f_Y denotes the population distribution of Y:

$$\mathcal{S}_{\delta,\epsilon} = \left\{\delta_{\theta_0} \,:\, f_{\theta_0} = f_Y,\, \|\theta_0 - \theta^*\|_\Omega^2 \leq \epsilon\right\}.$$

$\mathcal{S}_{\delta,\epsilon}$ is the identified set for $\delta(\theta_0)$, subject to the restriction that $\|\theta_0 - \theta^*\|_\Omega^2 \leq \epsilon$.

The first part in the following result shows that, for vanishing ϵ, $\mathcal{S}_{\delta,\epsilon}$ is contained in $\mathbb{E}_Y(h(Y)) \pm b_\epsilon(h, \theta^*)$, for all h satisfying (1). The second part shows that $\mathcal{S}_{\delta,\epsilon}$ actually coincides with $\mathbb{E}_Y(\overline{h}(Y)) \pm b_\epsilon(\overline{h}, \theta^*)$ for a suitable moment function $\overline{h}$ satisfying (1).

Proposition 3

(i) Let $\delta_{\theta_0} \in \mathcal{S}_{\delta,\epsilon}$. Let h satisfying (1). Then, as ϵ tends to zero:

$$|\delta_{\theta_0} - \mathbb{E}_Y(h(Y))| \leq b_\epsilon(h, \theta^*) + o(\epsilon^{\frac{1}{2}}).$$

(ii) Moreover, there exists a moment function $\overline{h}$ satisfying (1) such that $\mathcal{S}_{\delta,\epsilon}$ coincides with the interval $\mathbb{E}_Y(\overline{h}(Y)) \pm b_\epsilon(\overline{h}, \theta^*)$, up to some $o(\epsilon^{\frac{1}{2}})$ terms.

9 Conclusion

To be completed.

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

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] 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.

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

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

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

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

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

[17] Fessler, P., and M. Kasy (2017): "How to Use Economic Theory to Improve Estimators," unpublished manuscript.

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

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

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

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

[22] 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.

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

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

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

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

[27] 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.

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

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

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

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

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

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

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

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

[36] 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.

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

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

[39] Rieder, H. (1994): *Robust Asymptotic Statistics*.

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

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

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

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

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

APPENDIX

A Proofs

Proof of Corollary 1. Note first that the Kullback-Leibler divergence is locally quadratic in the following sense:

$$
\begin{aligned}
d_{KL}(g_0\pi_0 f_X, g\pi f_X) &= 2\iiint \ln\left(\frac{g_0(y\,|\,\alpha,x)\pi_0(\alpha\,|\,x)f_X(x)}{g(y\,|\,\alpha,x)\pi(\alpha\,|\,x)f_X(x)}\right) g_0(y\,|\,\alpha,x)\pi_0(\alpha\,|\,x)f_X(x)dyd\alpha dx \\
&= 2\iiint \ln\left(\frac{g_0(y\,|\,\alpha,x)}{g(y\,|\,\alpha,x)}\right) g_0(y\,|\,\alpha,x)\pi_0(\alpha\,|\,x)f_X(x)dyd\alpha dx \\
&\quad + 2\iint \ln\left(\frac{\pi_0(\alpha\,|\,x)}{\pi(\alpha\,|\,x)}\right)\pi_0(\alpha\,|\,x)f_X(x)d\alpha dx \\
&= \iiint \frac{(g_0(y\,|\,\alpha,x)-g(y\,|\,\alpha,x))^2}{g(y\,|\,\alpha,x)}\pi_0(\alpha\,|\,x)f_X(x)dyd\alpha dx \\
&\quad + \iint \frac{(\pi_0(\alpha\,|\,x)-\pi(\alpha\,|\,x))^2}{\pi(\alpha\,|\,x)}f_X(x)d\alpha dx + o(\epsilon),
\end{aligned}
$$

as ϵ tends to zero, for g_0, π_0, g, π, f_X such that $d_{KL}(g_0\pi_0 f_X, g\pi f_X) \leq \epsilon$.

The Lagrangian associated with the maximization of:

$$
\left|\delta(g^*,\pi_0) - \iint h(y,x) f_{g^*,\pi_0}(y\,|\,x)f_X(x)dydx\right|
$$

is:

$$
\begin{aligned}
\mathcal{L} &= \iiint \left(\Delta(y,\alpha,x) - h(y,x)\right) g^*(y\,|\,\alpha,x)\pi_0(\alpha\,|\,x)f_X(x)dyd\alpha dx \\
&\quad + \int \lambda_1(x)\int \pi_0(\alpha\,|\,x)f_X(x)d\alpha dx + \lambda_2 \iint \frac{(\pi_0(\alpha\,|\,x)-\pi^*(\alpha\,|\,x))^2}{\pi^*(\alpha\,|\,x)}f_X(x)d\alpha dx,
\end{aligned}
$$

where $\lambda_1(\cdot), \lambda_2$ are associated with the integral constraint and local Kullback-Leibler constraint, respectively.

The first-order conditions with respect to π_0 are then:

$$
\begin{aligned}
&\int \left(\Delta(y,\alpha,x) - h(y,x)\right) g^*(y\,|\,\alpha,x)dy f_X(x) + \lambda_1(x)f_X(x) \\
&\qquad + 2\lambda_2 \frac{\pi_0(\alpha\,|\,x)-\pi^*(\alpha\,|\,x)}{\pi^*(\alpha\,|\,x)}f_X(x) = 0.
\end{aligned}
$$

It follows from integrating this equation with respect to $\pi^*(\alpha\,|\,x)$ that $\lambda_1(x) = 0$. Hence:

$$
\pi_0(\alpha\,|\,x) - \pi^*(\alpha\,|\,x) = -\frac{1}{2\lambda_2}\pi^*(\alpha\,|\,x)\mathbb{E}_*\left(\Delta(Y,\alpha,x) - h(Y,x)\,|\,\alpha,x\right).
$$

Using the Kullback-Leibler constraint gives:

$$\pi_0(\alpha \,|\, x) - \pi^*(\alpha \,|\, x) = \pm\epsilon^{\frac{1}{2}} \frac{\pi^*(\alpha \,|\, x)\mathbb{E}_* \left(\Delta(Y,\alpha,x) - h(Y,x) \,|\, \alpha, x\right)}{\left\{ \mathrm{Var}_* \left[\mathbb{E}_* \left(\Delta(Y,\alpha,x) - h(Y,x) \,|\, \alpha, x\right) \right] \right\}^{\frac{1}{2}}}. \tag{A1}$$

The first part in Corollary 1 follows directly.

The minimum-sensitivity moment function is obtained by minimizing $b_\epsilon(g^*, \pi^*, h)$ subject to (18). The associated Lagrangian is:

$$\begin{aligned}\mathcal{L} = &\iint \left[\int \left(\Delta(y,\alpha,x) - h(y,x)\right) g^*(y \,|\, \alpha, x) dy \right]^2 \pi^*(\alpha \,|\, x) f_X(x) d\alpha dx \\ &+ \int \lambda(x) \left[\iint h(y,x) g^*(y \,|\, \alpha, x) \pi^*(\alpha \,|\, x) dy d\alpha \right] f_X(x) dx,\end{aligned}$$

the first-order conditions of which are (divided by $f_X(x)$):

$$\begin{aligned}&- \int g^*(y \,|\, \alpha, x) \left[\int \left(\Delta(y,\alpha,x) - h^{MS}_{g^*,\pi^*}(y,x)\right) g^*(y \,|\, \alpha, x) dy \right] \pi^*(\alpha \,|\, x) d\alpha \\ &\quad + \lambda(x) \int g^*(y \,|\, \alpha, x) \pi^*(\alpha \,|\, x) d\alpha = 0.\end{aligned}$$

Integrating this expression with respect to y gives $\lambda(x) = 0$ for all x. Finally, dividing by $\int g^*(y \,|\, \alpha, x)\pi^*(\alpha \,|\, x) d\alpha$, and noting that $\frac{g^*(y \,|\, \alpha, x)\pi^*(\alpha \,|\, x)}{\int g^*(y \,|\, a, x)\pi^*(a \,|\, x) da} = p^*(\alpha \,|\, y, x)$ is the posterior density of α given (y, x), shows the second part of Corollary 1.

Proof of Corollary 2. In this case the Lagrangian associated with the maximization of $\left|\delta(g^*, \pi_0) - \iint h(y,x) f_{g^*,\pi_0}(y \,|\, x) f_X(x) dy dx\right|$ is:

$$\begin{aligned}\mathcal{L} = &\iiint \left(\Delta(y,\alpha,x) - h(y,x)\right) g^*(y \,|\, \alpha, x) \pi_0(\alpha \,|\, x) f_X(x) dy d\alpha dx \\ &+ \int \lambda_1(x) \int \pi_0(\alpha \,|\, x) f_X(x) d\alpha dx + 2\lambda_2 \iint \ln\left(\frac{\pi_0(\alpha \,|\, x)}{\pi^*(\alpha \,|\, x)}\right) \pi_0(\alpha \,|\, x) f_X(x) d\alpha dx.\end{aligned}$$

The first-order conditions with respect to π_0 are then (dividing by $f_X(x)$):

$$\int \left(\Delta(y,\alpha,x) - h(y,x)\right) g^*(y \,|\, \alpha, x) dy + \left[\lambda_1(x) + 2\lambda_2\right] + 2\lambda_2 \ln\left(\frac{\pi_0(\alpha \,|\, x)}{\pi^*(\alpha \,|\, x)}\right) = 0.$$

Hence, using that $\pi_0(\alpha \,|\, x)$ integrates to one for all x:

$$\pi_0(\alpha \,|\, x) = C\pi^*(\alpha \,|\, x) \exp\left(-\frac{1}{2\lambda_2} \int \left(\Delta(y,\alpha,x) - h(y,x)\right) g^*(y \,|\, \alpha, x) dy \right),$$

for:

$$C^{-1} = \int \pi^*(a \,|\, x) \exp\left(-\frac{1}{2\lambda_2} \int \left(\Delta(y,a,x) - h(y,x)\right) g^*(y \,|\, a, x) dy \right) da. \tag{A2}$$

Since, at the least-favorable π_0, $2\iint \ln\left(\frac{\pi_0(\alpha\,|\,x)}{\pi^*(\alpha\,|\,x)}\right)\pi_0(\alpha\,|\,x)f_X(x)d\alpha dx = \epsilon$, we have:

$$\epsilon = 2\ln C + 2C\iint\left(-\frac{1}{2\lambda_2}\int\left(\Delta(y,\alpha,x)-h(y,x)\right)g^*(y\,|\,\alpha,x)dy\right) \times \pi^*(\alpha\,|\,x)\exp\left(-\frac{1}{2\lambda_2}\int\left(\Delta(y,\alpha,x)-h(y,x)\right)g^*(y\,|\,\alpha,x)dy\right)f_X(x)d\alpha dx. \tag{A3}$$

It follows that:

$$b_\epsilon(g^*,\pi^*,h) = \epsilon^{\frac{1}{2}}\Bigg\{C\iint\left(\int\left(\Delta(y,\alpha,x)-h(y,x)\right)g^*(y\,|\,\alpha,x)dy\right) \times \exp\left(-\frac{1}{2\lambda_2}\int\left(\Delta(y,\alpha,x)-h(y,x)\right)g^*(y\,|\,\alpha,x)dy\right)\pi^*(\alpha\,|\,x)f_X(x)d\alpha dx\Bigg\}^{\frac{1}{2}},$$

where C and λ_2 satisfy (A2)-(A3).

Proof of Corollary 3. Let $f_0 = g_0\pi_0$. The Lagrangian is:

$$\begin{aligned}\mathcal{L} = &\iiint\left(\Delta(y,\alpha,x)-h(y,x)\right)f_0(y,\alpha\,|\,x)f_X(x)dyd\alpha dx \\ &+\int\lambda_1(x)\left(\iint f_0(y,\alpha\,|\,x)dyd\alpha\right)f_X(x)dx \\ &+\lambda_2\iiint\frac{\left(f_0(y,\alpha\,|\,x)-g^*(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x)\right)^2}{g^*(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x)}f_X(x)dyd\alpha dx,\end{aligned}$$

and the first-order conditions with respect to f_0 are (divided by $f_X(x)$):

$$\left(\Delta(y,\alpha,x)-h(y,x)\right)+\lambda_1(x)+2\lambda_2\frac{f_0(y,\alpha\,|\,x)-g^*(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x)}{g^*(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x)} = 0.$$

It is easy to see that $\lambda_1(x) = 0$. Moreover:

$$f_0(y,\alpha\,|\,x)-g^*(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x) = \pm\epsilon^{\frac{1}{2}}\frac{g^*(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x)\left(\Delta(y,\alpha,x)-h(y,x)\right)}{\left\{\mathrm{Var}_*\left[\Delta(Y,A,X)-h(Y,X)\right]\right\}^{\frac{1}{2}}}.$$

This gives the form of $b_\epsilon(g^*,\pi^*,h)$. The form for $h^{MS}_{g^*,\pi^*}$ comes from the characterization of the conditional expectation as the best predictor under square loss.

Proof of Corollary 4. In this case the Lagrangian is:

$$\begin{aligned}\mathcal{L} = &\iiint\left(\Delta(y,\alpha,x)-h(y,x)\right)g_0(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x)f_X(x)dyd\alpha dx \\ &+\iint\lambda_1(\alpha,x)\left(\int g_0(y\,|\,\alpha,x)dy\right)\pi^*(\alpha\,|\,x)f_X(x)d\alpha dx \\ &+\lambda_2\iiint\frac{\left(g_0(y\,|\,\alpha,x)-g^*(y\,|\,\alpha,x)\right)^2}{g^*(y\,|\,\alpha,x)}\pi^*(\alpha\,|\,x)f_X(x)dyd\alpha dx,\end{aligned}$$

and the first-order conditions with respect to g_0 are (divided by $f_X(x)$):

$$\begin{aligned}&\left(\Delta(y,\alpha,x)-h(y,x)\right)\pi^*(\alpha\,|\,x)\\&\quad+\lambda_1(\alpha,x)\pi^*(\alpha\,|\,x)+2\lambda_2\frac{g_0(y\,|\,\alpha,x)-g^*(y\,|\,\alpha,x)}{g^*(y\,|\,\alpha,x)}\pi^*(\alpha\,|\,x)=0.\end{aligned}$$

Similarly as in the proof of Corollary 1 we obtain:

$$\lambda_1(\alpha,x)=-\mathbb{E}_*\left[\Delta(Y,\alpha,x)-h(Y,x)\,|\,\alpha,x\right],$$

and:

$$\begin{aligned}&g_0(y\,|\,\alpha,x)-g^*(y\,|\,\alpha,x)\\&\quad=\pm\epsilon^{\frac{1}{2}}\frac{g^*(y\,|\,\alpha,x)\left(\Delta(y,\alpha,x)-h(y,x)-\mathbb{E}_*\left[\Delta(Y,\alpha,x)-h(Y,x)\,|\,\alpha,x\right]\right)}{\left\{\mathrm{Var}_*\left[\Delta(Y,A,X)-h(Y,X)-\mathbb{E}\left[\Delta(\widetilde{Y},A,X)-h(\widetilde{Y},X)\,|\,A,X\right]\right]\right\}^{\frac{1}{2}}}.\end{aligned}$$

The end of the proof follows similar steps as in the proof of Corollary 1.

Proof of Corollary 5. For small ϵ, minimizing the MSE amounts (up to lower-order terms) to minimizing the following quantity with respect to h:

$$\begin{aligned}\epsilon\iint\left[\mathbb{E}_*\left(\Delta(Y,\alpha,x)-h(Y,x)\,|\,\alpha,x\right)\right]^2\pi^*(\alpha\,|\,x)f_X(x)d\alpha dx\\+\frac{1}{N}\iint h^2(y,x)f_{g^*,\pi^*}(y,x)dydx,\end{aligned}$$

subject to:

$$\iint\left(\Delta(y,\alpha,x)-h(y,x)\right)g^*(y\,|\,\alpha,x)\pi^*(\alpha\,|\,x)dyd\alpha=0,\ \text{for all } x.$$

Solving this system using variational calculus implies Corollary 5.

Proof of Proposition 3. Let us start with Part (i). Let $\delta_{\theta_0}\in\mathcal{S}_{\delta,\epsilon}$, and h such that (1) holds. We have:

$$\begin{aligned}\left|\delta_{\theta_0}-\mathbb{E}_Y(h(Y))\right|&=\left|\delta_{\theta_0}-\int h(y)f_Y(y)dy\right|\\&=\left|\delta_{\theta_0}-\int h(y)f_{\theta_0}(y)dy\right|\\&=\left|\left(\nabla\delta_{\theta^*}-\int h(y)\nabla f_{\theta^*}(y)dy\right)(\theta_0-\theta^*)\right|+o(\epsilon^{\frac{1}{2}})\\&\leq b_\epsilon(h,\theta^*)+o(\epsilon^{\frac{1}{2}}).\end{aligned}$$

Let us now turn to Part (ii). The Lagrangian associated with the upper/lower bound of $\mathcal{S}_{\delta,\epsilon}$ is:

$$\mathcal{L} = \delta_{\theta_0} + \int \lambda_1(y) f_{\theta_0}(y) dy + \lambda_2 \|\theta_0 - \theta^*\|_\Omega^2,$$

and the first-order conditions with respect to θ_0 give:

$$\begin{aligned}\theta_0 - \theta^* &= -\frac{1}{2\lambda_2}\Omega^{-1}\left(\nabla\delta_{\theta^*} + \int \lambda_1(y)\nabla f_{\theta^*}(y)dy\right) + o(\epsilon^{\frac{1}{2}}) \\ &= \pm\epsilon^{\frac{1}{2}} \frac{\Omega^{-1}\left(\nabla\delta_{\theta^*} + \int \lambda_1(y)\nabla f_{\theta^*}(y)dy\right)}{\left\|\nabla\delta_{\theta^*} + \int \lambda_1(y)\nabla f_{\theta^*}(y)dy\right\|_{\Omega^{-1}}} + o(\epsilon^{\frac{1}{2}}).\end{aligned}$$

Let:

$$\overline{h}(y) \equiv \delta_{\theta^*} + \int \lambda_1(y) f_{\theta^*}(y)dy - \lambda_1(y).$$

Notice that $\overline{h}$ satisfies (1). Moreover:

$$\begin{aligned}|\delta_{\theta_0} - \mathbb{E}_Y(\overline{h}(Y))| &= \left|\delta_{\theta^*} + \nabla\delta_{\theta^*}(\theta_0 - \theta^*) - \int \overline{h}(y) f_{\theta_0}(y)dy\right| + o(\epsilon^{\frac{1}{2}}) \\ &= \left|\delta_{\theta^*} - \int \overline{h}(y) f_{\theta^*}(y)dy + \left(\nabla\delta_{\theta^*} - \int \overline{h}(y)\nabla f_{\theta^*}(y)dy\right)(\theta_0 - \theta^*)\right| + o(\epsilon^{\frac{1}{2}}) \\ &= \epsilon^{\frac{1}{2}}\left|\frac{\left(\nabla\delta_{\theta^*} - \int \overline{h}(y)\nabla f_{\theta^*}(y)dy\right)'\Omega^{-1}\left(\nabla\delta_{\theta^*} + \int \lambda_1(y)\nabla f_{\theta^*}(y)dy\right)}{\left\|\nabla\delta_{\theta^*} + \int \lambda_1(y)\nabla f_{\theta^*}(y)dy\right\|_{\Omega^{-1}}}\right| \\ &= \epsilon^{\frac{1}{2}}\left\|\nabla\delta_{\theta^*} - \int \overline{h}(y)\nabla f_{\theta^*}(y)dy\right\|_{\Omega^{-1}} + o(\epsilon^{\frac{1}{2}}) = b_\epsilon(\overline{h}, \theta^*) + o(\epsilon^{\frac{1}{2}}),\end{aligned}$$

where we have used that:

$$\int \overline{h}(y)\nabla f_{\theta^*}(y)dy = -\int \lambda_1(y)\nabla f_{\theta^*}(y)dy.$$