Robustness to Missing Data: Breakdown Point Analysis

Daniel Ober-Reynolds[*]

This draft: 21 June, 2023
First draft: 12 January, 2023

Abstract

Missing data is pervasive in econometric applications, and rarely is it plausible that the data are missing (completely) at random. This paper proposes a methodology for studying the robustness of results drawn from incomplete datasets. Selection is measured as the squared Hellinger divergence between the distributions of complete and incomplete observations, which has a natural interpretation. The *breakdown point* is defined as the minimal amount of selection needed to overturn a given result. Reporting point estimates and lower confidence intervals of the breakdown point is a simple, concise way to communicate a result's robustness. An estimator of the breakdown point of results drawn from GMM models is proposed and shown $\sqrt{n}$-consistent and asymptotically normal under mild assumptions. Lower confidence intervals of the breakdown point are constructed with a simple bootstrap procedure. The paper concludes with a simulation study illustrating the good finite sample performance of the procedure.

[*] doberreynolds [at] gmail.com. I want to thank Andres Santos for his continuous guidance and support.

1 Introduction

Virtually every economic dataset is plagued by missing and incomplete records. Survey nonresponse is the most visible cause, and appears to be worsening over time. Bollinger et al. (2019) report that the Current Population Survey's Annual Social and Economics Supplement item and whole nonresponse has been increasing, reaching 43% in 2015. By linking these data with the Social Security Administration Detailed Earnings Record, the authors show that the distribution of nonreponders differs from that of responders even after conditioning on a large set of covariates.

Missing data is not an issue likely to be solved by new technology or collection methods. Privacy concerns around the use of smart phones and social media to collect user data are growing, and companies and policy makers are responding forcefully. Two notable examples are Apple's "App Tracking Transparency" feature, which allows users to opt-out of data sharing with a simple button, and the European Union's General Data Protection Regulation, which requires companies obtain customer consent "by a statement or by a clear affirmative action" (European Parliament and Council of the European Union (2016), Article 4.11). Both policies appear to have bite. The former reportedly cost Meta 10 billion dollars in ad revenue in 2022 (Isaac (2022)), and Google was fined 50 million euros in 2019 for violating the latter (Satariano (2019)).

Samples with missing or incomplete observations fail to identify the population distribution (Manski (2005)). To make progress, researchers commonly apply standard procedures to the complete observations. This practice is typically justified by assuming the data are "missing completely at random" (MCAR); that is, incomplete observations follow the same distribution as that of the complete observations. In many settings such an assumption is implausible. Without it, the conclusions drawn are uncomfortably qualified as being about the distribution of the complete observations, rather than the actual distribution of interest.

This paper proposes a method to investigate the robustness of a conclusion when asserted about the whole population. Results are more robust when overturning them would require more selection. To make this intuition precise, selection is measured with the squared Hellinger divergence between the distribution of complete observations and that of the incomplete observations. Although many different statistical divergences could be used to measure selection, squared Hellinger is interpretable as a measure of how well the variables under study would predict an observation being complete. This gives the values of the selection measure context, allowing researchers to guage how much selection can be expected in a given setting.

The *breakdown point* is the minimum amount of selection needed to overturn a conclusion. Readers who doubt the setting exhibits that much selection will find the conclusion compelling. In models identified with the generalized method of moments (GMM), the breakdown point is the constrained minimum of the value function of a convex optimization problem. Estimators of the breakdown point are constructed from the dual of this convex inner problem, and shown to be $\sqrt{n}$-consistent and asymptotically normal. Lower confidence intervals are simple to construct. Reporting the point estimates and lower confidence intervals of the breakdown point is a simple, concise way to communicate a result's robustness.

This approach has a number of advantages over existing methods for incomplete datasets. Sample selection models consider regressions with samples where the dependent variable is sometimes missing, and obtain point identification by modeling the selection process (Heckman (1979), Das et al. (2003)). These models require the data include a variable changing the probability of observation but not the dependent variable. This "exclusion restriction" is difficult to satisfy in many applications. The breakdown point approach proposed here can be used on most common GMM models (including but not restricted to regressions with missing outcomes), and requires no additional data. The breakdown point can be estimated even if the incomplete observations are in fact completely missing, a distinct possibility when using survey data.

The econometric literature on missing data has also explored bounding the parameter of interest based on the support (Manski (2005), Horowitz & Manski (2006)). If all parameter values within these "worst-case" bounds satisfy the researcher's conclusion, then the conclusion is undoubtedly robust. Unfortunately, the bounds may be uninformative in practice. Proponents of this approach are well aware these bounds are conservative, and propose this exercise as a place to begin an investigation rather than end one. Additional identifying assumptions should then be considered, in order to make plain to readers what needs to be assumed to reach a given conclusion. (See, e.g., Manski (2013) section 3.) The breakdown approach proposed here is a simple version of this type of exercise, as the assumption that selection is less than the breakdown point leads one to conclude the hypothesis under investigation.

A growing literature advocates for breakdown analysis as a general, tractable method to assess the sensitivity of a result to relaxations of identifying assumptions. The term "identification breakdown point" can be found as early as Horowitz & Manski (1995) in the context of corrupted data. Masten & Poirier (2020) advocates for the approach generally, and illustrates it with the potential outcomes framework. Diegert et al. (2022) define and study breakdown points in linear regressions suffering from omitted variable bias.

This paper is not the first to notice the appeal of breakdown point analysis in the context of missing data. Kline & Santos (2013) consider a setting with a missing scalar, propose measuring selection with the maximal Kolmogorov-Smirnov (KS) distance between the conditional distributions of complete and incomplete observations across all values of covariates, and advocate for "reporting the minimal level of selection necessary to undermine a hypothesis," (p. 233). The methodology proposed here has some notable advantages. First, measuring selection with the maximal KS distance limits researchers to the case where only a scalar is missing, while measuring selection as proposed here allows any number of variables to be missing. Second, in a given setting it is easier to gauge whether the variables under study are likely to be good predictors of missingness than what share of the missing data is missing at random. This makes squared Hellinger a more natural measure of selection than KS distance. Which approach is more tractable will depend on the parameter of interest. Kline & Santos (2013) derives sharp, closed form bounds to the conditional quantiles of the missing variable, and frame the conclusion to be investigated in terms of those quantiles. This paper assumes the parameter of interest is identified with GMM and uses the model directly, giving up closed form solutions. In thoery this could lead to computational difficulties for complex models, but the simulations in section 5 present no issue.

The remainder of this paper is structured as follows: section 2 formalizes the setting, the proposed measure of selection, and the breakdown point. The dual problem is presented and discussed in section 3. Section 4 defines the estimator and states the main results on estimation and inference, which are proven in the appendix. Section 5 presents a simulation study investigating the finite sample performance of these estimators. Section 6 concludes.

2 Measuring selection and breakdown analysis

Suppose the available data is the i.i.d. sample $\{(D_i, D_iY_i, X_i)\}_{i=1}^n$, where $Z_i = (Y_i, X_i) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ contains the variables of interest and $D_i \in \{0,1\}$ indicates whether Y_i is observed. Note that Y_i may be a vector, and X_i may be empty. Let $p_D = P(D=1)$ denote the probability of observing Y, P_1 the distribution of $Z = (Y, X)$ conditional on $D = 1$, and P_0 the distribution of Z conditional on $D = 0$. P_1 and P_0 are called the *complete case* and *incomplete case* distributions respectively. The distribution of interest is the unconditional distribution of Z, given by $p_D P_1 + (1-p_D)P_0$. When X is nonempty, the marginal distribution of X conditional on $D = 0$ is denoted P_{0X}. For simplicity,

X is assumed to have the same finite support when $D = 0$ as when $D = 1$, which greatly simplifies asymptotic analysis. Remark 2.3 discusses this assumption further.

To fix ideas, consider data collection via survey. Y is a vector of data the survey hopes to collect, which is observed only if the recipient responds ($D = 1$). The survey's response rate, $p_D = P(D = 1)$, is essentially always less than one in practice. It is common for administrative data to provide basic information about a survey recipient (such as age, occupation, etc.), which is collected in X.

Analyses based on the complete observations may not convince researchers who worry that P_0 differs from P_1. Such concerns are common, as few settings plausibly satisfy the missing completely at random assumption. However, it is often similarly implausible that P_0 differs greatly from P_1. Researchers who convincingly argue that P_0 is not too different from P_1 can still convince their audience of conclusions drawn from an analysis of P_1.[1]

A quantitative measure of the difference between P_1 and P_0 is needed to make this argument formal and convincing. The statistics literature provides a natural solution in the form of *divergences*: functions mapping two probability distributions to the nonnegative real line that take value zero if and only if the two distributions are the same. There are many such functions. To be useful as a measure of selection, a divergence should have a tractable interpretation, so that researchers can gauge whether a given amount of selection is reasonable for their setting.

2.1 An interpretable measure of selection

Missing data cause greater concern when researchers expect the variables of interest (Z) to be a good predictor of incompleteness (D). Consider again the example of data collection via survey. Researchers are rightfully more concerned about survey nonresponse when asking about the respondent's arrest record than when asking for opinions on recent television programming. People with criminal records may be less willing to answer questions about that record.[2] This suggests that the distribution of responders may look quite different from the distribution of nonresponders, and that criminal records would be a good predictor of nonresponse.

To illustrate this more formally, let f_1 and f_0 be densities of P_1 and P_0 with respect to $p_D P_1 +$

[1] In some cases, such as correctly specified regression models, it suffices that the conditional distributions $f_{Y|X=x,D=0}(y \mid x)$ are the same as the identified $f_{Y|X=x,D=1}(y \mid x)$. This weaker "missing at random" (MAR) assumption is also rarely plausible in practice, and analyses based on this assumption often rely heavily on the model being correctly specified.

[2] For example, Brame et al. (2012) estimate the cumulative prevalence of arrest from ages 8 to 23 from a survey directly asking about prior arrests. The authors report upper and lower bounds derived by assuming the entire set of nonresponders had or had not been arrested, essentially the worst-case bounds advocated for by Manski (2005).

$(1 - p_D)P_0$ respectively:

$$f_1(z) = \frac{P(D = 1 \mid Z = z)}{p_D}, \qquad\qquad f_0(z) = \frac{(1 - P(D = 1 \mid Z = z))}{1 - p_D}$$

An optimist may assume D is independent of Z, implying that $P(D = 1 \mid Z = z) = P(D = 1)$ and $f_0 = f_1 = 1$. In contrast, a pessimist may assume D is close to a deterministic function of Z, allowing Z to predict D well. This would imply $P(D = 1 \mid Z = z)$ is close to 1 or 0 for many values of z, and thus f_1 differs greatly from f_0.

As in the survey example, the setting often makes it clear whether Z would be a good predictor of D. This hueristic is useful to identify and discuss selection concerns. The following lemma shows that measuring selection as the squared Hellinger distance between P_0 and P_1 captures this intuition, with larger values corresponding to Z having greater capability of predicting D.[3]

Lemma 2.1. *Let $(Z, D) \in \mathbb{R}^{d_z} \times \{0, 1\}$ be random variables with $p_D = P(D = 1) \in (0, 1)$. Let $Z \mid D = 1 \sim P_1$ and $Z \mid D = 0 \sim P_0$. Then*

$$H^2(P_0, P_1) = 1 - \frac{E\left[\sqrt{Var(D \mid Z)}\right]}{\sqrt{Var(D)}} \tag{1}$$

where the expectation is taken with respect to $p_D P_1 + (1 - p_D)P_0$, the marginal distribution of Z.

All results are proven in the appendix.

Equation (1) states that (one half times) the squared Hellinger distance between P_0 and P_1 is the expected percent of the standard deviation of D reduced by conditioning on Z. In the extreme case where $\text{Var}(D \mid Z) = \text{Var}(D)$, equation (1) implies $H^2(P_0, P_1) = 0$ and the conditional distributions are the same. As the ability of Z to predict D grows, the variance of D conditional on Z decreases and $H^2(P_0, P_1)$ grows toward one.

Remark 2.1. It's straightforward to see that $\text{Var}(D \mid X) \geq \text{Var}(D \mid X, Y)$ implies

$$H^2(P_0, P_1) = 1 - \frac{E\left[\sqrt{\text{Var}(D \mid Y, X)}\right]}{\sqrt{\text{Var}(D)}} \geq 1 - \frac{E[\sqrt{\text{Var}(D \mid X)}]}{\sqrt{\text{Var}(D)}} = H^2(P_{0X}, P_{1X})$$

[3] The Hellinger distance between probability measures Q and P is

$$H(Q, P) = \left(\frac{1}{2}\int\left(\sqrt{\frac{dQ}{d\lambda}(z)} - \sqrt{\frac{dP}{d\lambda}(z)}\right)^2 d\lambda(z)\right)^{1/2}$$

where λ is any measure dominating both P and Q.

where P_{0X}, P_{1X} are the marginal distributions of X conditonal on $D = 0$ and $D = 1$ respectively. This lower bound on the selection is identified from the sample, and motivates the common practice of comparing the distribution of X conditional on $D = 0$ with that of X conditional on $D = 1$; the distributions P_0 and P_1 can only be "further" apart.

2.2 Divergences

Squared Hellinger provides an intuitive measure of selection, but there are many other options. Recall that a function $d(\cdot\|\cdot)$ mapping two probability distributions P and Q to $\mathbb{R}$ is called a *divergence* if 1. $d(Q\|P) \geq 0$, and 2. $d(Q\|P) = 0$ if and only if $P = Q$. Divergences need not be symmetric nor satisfy the triangle inequality. The set of *f-divergences* are particularly well behaved. Given a convex function $f : \mathbb{R} \to [0, \infty]$ satisfying $f(t) = \infty$ for $t < 0$ and taking a unique minimum of $f(1) = 0$, the corresponding f-divergence is given by

$$d_f(Q\|P) = \begin{cases} \int f\left(\frac{dQ}{dP}\right) dP & \text{if } Q \ll P \\ \infty & \text{otherwise} \end{cases} \tag{2}$$

Many popular divergences are equal to f-divergences when P dominates Q.

Name	Common formula	$f(t)$ when $t \geq 0$
Squared Hellinger	$H^2(Q, P) = \frac{1}{2}\int\left(\sqrt{\frac{dQ}{dP}(z)} - 1\right)^2 dP(z)$	$f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$
Kullback-Leibler (KL)	$KL(Q\|P) = \int \log\left(\frac{dQ}{dP}(z)\right) dQ(z)$	$f(t) = t\log(t) - t + 1$
"Reverse" KL	$KL(P\|Q) = \int \log\left(\frac{dP}{dQ}(z)\right) dP(z)$	$f(x) = -\log(t) + t - 1$
Cressie-Read	–	$f_\gamma(t) = \frac{t^\gamma - \gamma t + \gamma - 1}{\gamma(\gamma - 1)}$, $\gamma < 1$

Table 1: Common f-divergences

Although squared Hellinger has intuitive appeal outlined in Section 2.1, the breakdown point analysis proposed in this paper remains tractable for any f-divergence listed in Table 1.[4] Precise assumptions regarding the f-divergence are collected in Assumption 1 below.

Remark 2.2. Measuring selection with an f-divergence facilitates estimation and inference, as the

[4] It is worth noting that the Cressie-Read divergence nests the other three as special cases. Squared Hellinger corresponds to $\frac{1}{2}f_{1/2}$. l'Hôpital's rule shows that Kullback-Leibler corresponds to $\lim_{\gamma\to 1} f_\gamma$ and Reverse Kullback-Leibler to $\lim_{\gamma\to 0} f_\gamma$. See Broniatowski & Keziou (2012) for additional discussion.

space of distributions Q with $d_f(Q\|P_1) < \infty$ corresponds to the set of densities with respect to P_1. In substance, this assumes $P_0 \ll P_1$ and rules out selection mechanisms that "truncate" data.

2.3 Breakdown analysis in GMM models

Suppose a preliminary analysis supports an alternative hypothesis H_1 over a null hypothesis H_0.[5] The breakdown point is the minimum amount of selection needed to overturn such a conclusion. When selection is measured in terms of the squared Hellinger distance, the breakdown point translates the claim that H_0 is true into a claim about the ability of Z to predict D. Specifically, if H_0 were true then $1 - \frac{E[\sqrt{\text{Var}(D|Z)}]}{\sqrt{\text{Var}(D)}}$ would be weakly larger than the breakdown point. If this claim is implausible, then H_0 is similarly implausible.

This section formalizes this idea for generalized method of moment (GMM) models. Suppose the parameter of interest $\beta \in \mathbf{B} \subseteq \mathbb{R}^{d_b}$ is characterized as the unique solution to a finite set of moment conditions,

$$E[g(Z, \beta)] = 0 \in \mathbb{R}^{d_g}$$

where the expectation is taken with respect to the unconditional distribution, $p_D P_1 + (1 - p_D) P_0$. The conclusion to be investigated is that β falls outside a particular set $\mathbf{B}_0 \subset \mathbf{B}$, motivating the null and alternative hypotheses

$$H_0 \ : \ \beta \in \mathbf{B}_0, \qquad\qquad H_1 \ : \ \beta \in \mathbf{B} \setminus \mathbf{B}_0$$

Recall that the observed data is $\{(D_i, D_i Y_i, X_i)\}_{i=1}^n$, where $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$. The sample identifies P_1, p_D, and P_{0X}. The *breakdown point* is the minimal amount of selection needed to *rationalize* p_D, P_1, P_{0X}, and $\beta \in \mathbf{B}_0$. A hypothetical distribution of the incomplete observations Q rationalizes the parameter b if it has the identified marginal distribution of X, $Q_X = P_{0X}$, and the implied unconditional distribution $p_D P_1 + (1 - p_D) Q$ solves the moment conditions for b. The set of such distributions is implying finite selection is

$$\mathbf{P}^b = \{Q \ ; \ Q \ll P_1, \ Q_X = P_{0X}, \ p_D E_{P_1}[g(Z, b)] + (1 - p_D) E_Q[g(Z, b)] = 0\}, \tag{3}$$

[5] For example, such an analysis may be based on the complete observations assuming MCAR, or using imputation and assuming Y is MAR conditional on X.

The breakdown point δ^{BP} is the minimum selection needed to rationalize the null hypothesis:

$$\delta^{BP} = \inf_{b \in \mathbf{B}_0} \inf_{Q \in \mathbf{P}^b} d_f(Q \| P_1) \tag{4}$$

where the infimum over the empty set is understood as ∞. A simple example illustrates the idea.

Example 2.1. *Let $Y \in \mathbb{R}$ and $\beta = E[Y] = p_D E_{P_1}[Y] + (1 - p_D) E_{P_0}[Y]$. Let $p_D = 0.7$ and P_1 be $\mathcal{U}[0,1]$. The claim to support is $H_1 : \beta > 0.4$, and selection is measured with squared Hellinger. $\boldsymbol{P}^b$ is the set of continuous distributions on $[0,1]$ with expectation $\frac{b - p_D/2}{1 - p_D}$, so that $Q \in \boldsymbol{P}^b$ implies*

$$p_D E_{P_1}[Y] + (1 - p_D) E_Q[Y] = \frac{p_D}{2} + (1 - p_D) \frac{b - p_D/2}{1 - p_D} = b$$

The inner minimization in (4) chooses the distribution that minimizes selection while rationalizing b. The outer minimization chooses the parameter that minimizes selection while rationalizing $H_0 : \beta \leq 0.4$. Unsurprisingly, the outer minimization is solved by $b = 0.4$. The breakdown point, δ^{BD}, is slightly above 0.2. A researcher convinced $H^2(P_0, P_1) \leq 0.2$ should conclude $\beta > 0.4$.

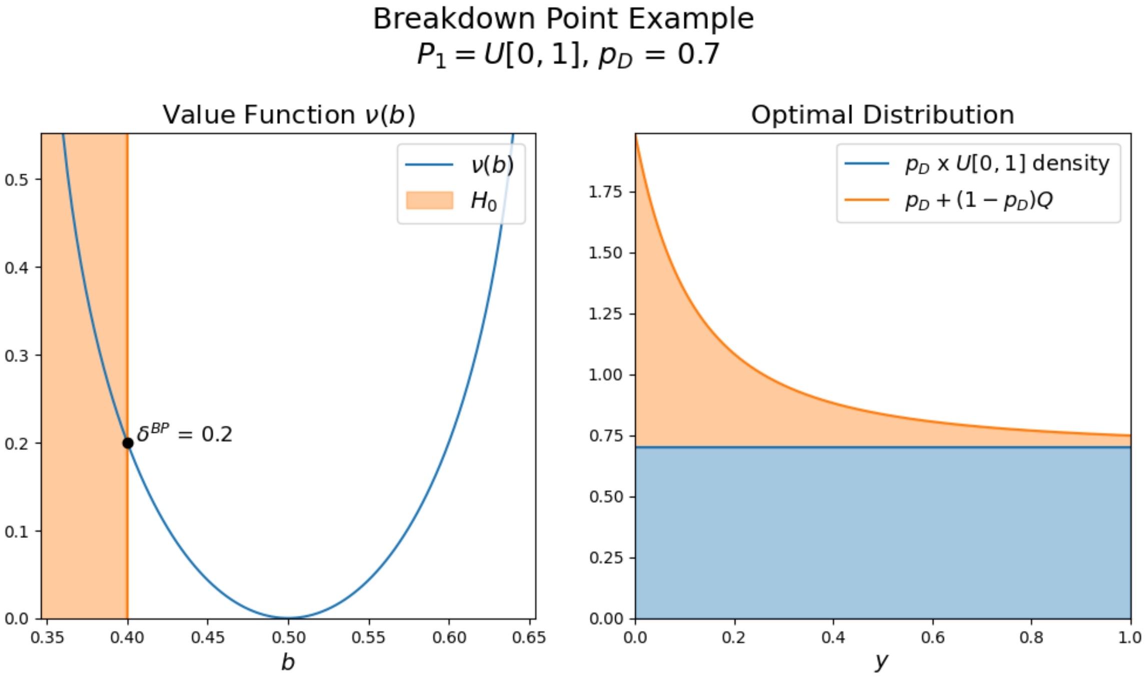


Figure 1: Left panel: $\nu(b)$, computed numerically using the dual problem discussed in section 3. Right panel: $p_D P_1$, and $p_D P_1 + (1 - pD) Q^*$, where $Q^* \in \mathbf{P}^{0.4}$ minimizes selection.

Breakdown analysis can also be framed as an exercise in partial identification, as in Kline & Santos (2013), Masten & Poirier (2020), and Diegert et al. (2022). In this framing, the researcher considers assumptions of the form $d_f(P_0, P_1) \leq \delta$ for some $\delta > 0$, which relax the assumption that

$P_0 = P_1$. The identified set for β grows with δ. As long as the identified set is a subset of $\mathbf{B} \setminus \mathbf{B}_0$, it is clear the researcher's conclusion holds. The breakdown point δ^{BP} can then be defined as either the largest δ for which the identified set is contained in $\mathbf{B} \setminus \mathbf{B}_0$, or the smallest δ for which the identified set has nontrivial intersection with $\mathbf{B}_0$ (the latter of which corresponds to the definition given in (4)). For further discussion of this equivalent framing of the breakdown point, see appendix B.2.

The remainder of this paper constructs a $\sqrt{n}$-consistent and asymptotically normal estimator of δ^{BP}, and provides a simple bootstrap procedure to construct consistent lower confidence intervals for δ^{BP}. Researchers working with partially complete datasets should discuss the plausible amount of selection in their setting, and report the point estimate and confidence interval lower bound for δ^{BP} for any conclusion of interest. This will make plain to readers which conclusions are more sensitive to missing data concerns, and whether crucial results are sufficiently robust.

2.4 Preview of results

The estimation proceeds by separating the optimizations in (4). Define the *primal problem*

$$\nu(b) = \inf_{Q \in \mathbf{P}^b} d_f(Q \| P_1) \tag{5}$$

and notice that $\delta^{BP} = \inf_{b \in \mathbf{B}_0} \nu(b)$. The first step is to estimate the value function ν over a set $B \subseteq \mathbf{B}$ large enough that $\inf_{b \in \mathbf{B}_0} \nu(b) = \inf_{b \in B \cap \mathbf{B}_0} \nu(b)$, while the second step estimates δ^{BP} through a simple plug-in estimator.

The primal problem is an infinite dimensional convex optimization problem over the space of probability distributions, but one that is very well studied in convex analysis. In particular, when $\mathbf{P}^b$ defined in (3) is characterized by a finite number of moment conditions, (5) has a well behaved, finite-dimensional dual problem with the same value function (Borwein & Lewis (1991), Borwein & Lewis (1993), Csiszár et al. (1999), Broniatowski & Keziou (2006)). Section 3 discusses this dual problem and the assumptions needed to make use of it. Under regularity conditions discussed in Section 4, sample analogue estimators based on this dual problem are uniformly consistent and asymptotically Gaussian on compact sets. Continuity and (Hadamard) differentiability of the infimum then implies consistency and convergence in distribution of the plug-in estimator.

To conclude this section, Assumption 1 collects conditions on the setting, the GMM model, and the f-divergence used to measure selection.

Assumption 1 (Setting)**.** *$\{(D_i, D_iY_i, X_i)\}_{i=1}^n$ is an i.i.d. sample from a distribution satisfying*

(i) $p_D = P(D = 1) \in (0, 1)$

(ii) $X \mid D = 1$ *and* $X \mid D = 0$ *have the same finite support* $\{x_1, \ldots, x_K\}$

(iii) $E\left[\sup_{b \in \boldsymbol{B}} \|g(Z, b)\| \mid D = 1\right] < \infty$*, where* $Z = (Y, X)$*,*

(iv) $f : \mathbb{R} \to [0, \infty]$ *is closed, proper, strictly convex, essentially smooth, takes its unique minimum of* $f(t) = 0$ *at* $t = 1$*, and satisfies* $f(t) = \infty$ *for all* $t < 0$*. The interior of* $dom(f) = \{t \in \mathbb{R}\ ;\ f(t) < \infty\}$*, denoted* (ℓ, u)*, satisfies* $\ell < 1 < u$*, and* f *is twice continuously differentiable on* (ℓ, u)*.*

The finite support condition in (ii) ensures that $\mathbf{P}^b$ defined in (3) is characterized by a finite number of moments (see remark 2.3 below for additional discussion). Condition (iv) ensures the f-divergence used to measure selection is well behaved, and is satisfied by every divergence in Table 1.[6] In particular, strict convexity of f ensures the primal problem (5) has a unique solution (P_1-almost surely). f is required to be essentially smooth to ensure the dual problem has a unique solution. The requirements that $f(x)$ take a unique minimum of 0 at $x = 1$ and $f(x) = \infty$ for $x < 0$ ensures that $d_f(Q\|P)$ is a well defined f-divergence.

Remark 2.3. If X is not finitely valued, it is easy to see that requiring Q_X match a finite number of moments of P_{0X} will estimate a point no larger than δ^{BP}. If this point is large enough to assuage missing data concerns, the breakdown point can only be larger. When the distribution of X is characterized by a countable set of moments, it may be possible to increase the number of moments with the sample size to estimate δ^{BP} directly. This is left for future research.

3 Duality

As defined in (5), $\nu(b)$ is the value function of an infinite dimensional convex optimization problem. Fortunately, when selection is measured with an f-divergence, (5) becomes a well-studied problem known by various names: maximal entropy (Csiszár et al. (1999)), partially finite programming (Borwein & Lewis (1991)), or simply f-divergence projection (Broniatowski & Keziou (2006)). The convex analysis results in these papers connect the primal problem in (5) to a finite dimensional dual problem that is much simpler to study and estimate. Under mild conditions, the value function of this dual problem coincides with the value function of the primal.

[6] See appendix C for definitions of the convex analysis terms used in Assumption 1 (iv).

To state the dual problem, first note that the primal can be viewed as a problem over the set of densities with respect to P_1:

$$\begin{aligned}\nu(b) = &\inf_q E\left[f(q(Y,X)) \mid D=1\right] \\ &\text{s.t. } E[h(Y,X,b)q(Y,X) \mid D=1] = c(b)\end{aligned}$$

where

$$h(z,b) = h(y,x,b) = \begin{pmatrix} g(y,x,b) \\ \mathbb{1}\{x = x_1\} \\ \vdots \\ \mathbb{1}\{x = x_K\}\end{pmatrix}, \qquad c(b) = \begin{pmatrix} \frac{-p_D}{1-p_D} E[g(Y,X,b) \mid D=1] \\ P(X = x_1 \mid D=0) \\ \vdots \\ P(X = x_K \mid D=0)\end{pmatrix}, \tag{6}$$

The dual problem corresponding to (5) is given by

$$V(b) = \sup_{\lambda \in \mathbb{R}^{d_g+K}} \lambda^\intercal c(b) - E\left[f^*\left(\lambda^\intercal h(Y,X,b)\right) \mid D=1\right] \tag{7}$$

where f^* is the convex conjugate of f, given by $f^*(r) = \sup_{t\in\mathbb{R}}\{rt - f(t)\}$.

Remark 3.1. To ensure q corresponds to a probability density, the constraints must enforce $\int q(z)dP(z) = 1$. This is implied by the constraints ensuring $Q_X = P_{0X}$ when X is present. If there are no always-observed variables, set $h(z,b) = \begin{pmatrix} g(z,b)^\intercal & 1\end{pmatrix}^\intercal \in \mathbb{R}^{d_g+1}$ and $c(b) = \begin{pmatrix}\frac{-p_D}{(1-p_D)}E[g(Y,X,b) \mid D=1]^\intercal & 1\end{pmatrix}^\intercal$.

For convenience, table 2 summarizes the convex conjugate for several common divergences. Just as the interior of dom(f) is denoted (ℓ, u), the interior of dom(f^*) will be denoted (ℓ^*, u^*).

Name	$f(t)$	ℓ, u	$f^*(r)$	ℓ^*, u^*
Squared Hellinger	$f(t) = \frac{1}{2}(\sqrt{t}-1)^2$	$\ell = 0,\ u = \infty$	$f^*(r) = \frac{1}{2}\left(\frac{1}{1-2r} - 1\right)$	$\ell^* = -\infty,\ u^* = 1/2$
Kullback-Leibler	$f(t) = t\log(t) - t + 1$	$\ell = 0,\ u = \infty$	$f^*(r) = \exp(r) - 1$	$\ell^* = -\infty,\ u^* = \infty$
"Reverse" KL	$f(t) = -\log(t) + t - 1$	$\ell = 0,\ u = \infty$	$f^*(r) = -\log(1-r)$	$\ell^* = -\infty,\ u^* = 1$

Table 2: Common f-divergence conjugates and effective domains

3.1 Weak and strong duality

Lemma D.2 in appendix D shows that under assumption 1, $V(b) \leq \nu(b)$. This fact is known as *weak duality*, and implies that

$$\inf_{b \in B \cap \mathbf{B}_0} V(b) \leq \inf_{b \in B \cap \mathbf{B}_0} \nu(b) = \delta^{BP} \tag{8}$$

for any $B \subseteq \mathbf{B}$. This inequality shows that using the dual problem for estimation of the breakdown point is at worst conservative: if $\inf_{b \in B \cap \mathbf{B}_0} V(b)$ is large enough to assuage selection concerns, researchers are assured that the breakdown point can only be larger.

Assuming only slightly more ensures *strong duality* holds, that is, $V(b) = \nu(b)$. Recall from Assumption 1 (iv) that the interior of $\text{dom}(f) = \{t \in \mathbb{R} \ ; \ f(t) < \infty\}$ is denoted (ℓ, u).

Assumption 2 (Strong duality)**.** *$B \subseteq \boldsymbol{B}$ is such that $\inf_{b \in \boldsymbol{B}_0} \nu(b) = \inf_{b \in B \cap \boldsymbol{B}_0} \nu(b)$. Furthermore, for each $b \in B$,*

(i) there exists $Q^b \in \boldsymbol{P}^b$ such that $\ell < \frac{\partial Q^b}{\partial P_1}(z) < u$, almost surely P_1

(ii) $\lambda(b)$ solving (7) is in the interior of $\{\lambda \ ; \ E[|f^*(\lambda^\intercal h(Z, b))| \mid D = 1] < \infty\}$

That strong duality holds under these conditions is a well known result.[7]

Theorem 3.1 (Strong duality)**.** *Suppose assumptions 1 and 2 hold. Then for each $b \in B$, $\nu(b) = V(b)$, with dual attainment.*

The first order condition of the dual problem (7) provides intuition. Exchanging expectation and differentiation, the first order condition is

$$\begin{pmatrix} \frac{-p_D}{1-p_D} E_{P_1}[g(Y, X, b)] \\ P(X = x_1 \mid D = 0) \\ \vdots \\ P(X = x_K \mid D = 0) \end{pmatrix} = E_{P_1}\left[(f^*)'\left(\lambda(b)^\intercal h(Y, X, b)\right) \begin{pmatrix} g(Y, X, b) \\ \mathbb{1}\{X = x_1\} \\ \vdots \\ \mathbb{1}\{X = x_K\} \end{pmatrix} \right]$$

where $\lambda(b) \in \mathbb{R}^{d_g + K}$ solves the dual problem. Consider $(f^*)'(\lambda(b)^\intercal h(y, x, b))$ as a density with respect to P_1. Notice that the first d_g equations of the first order condition ensure $p_D E_{P_1}[g(Y, X, b)] + (1 - p_D) E_{P_1}[(f^*)'\left(\lambda(b)^\intercal h(Y, X, b)\right) g(Y, X, b)] = 0$, while the remaining equalities ensure the marginal

[7] To the authors knowledge, the first to show strong duality holds rigorously was Borwein & Lewis (1991). The proof of theorem 3.1, found in appendix D, uses a result due to Csiszár et al. (1999).

distribution of X matches P_{0X}. In fact, the proof of theorem 3.1 shows that under assumptions 1 and 2, $(f^*)'\left(\lambda(b)^\intercal h(y,x,b)\right)$ is the P_1-density of the solution to the primal.

Assumption 2 ensures the set on which ν is estimated is large enough to estimate the breakdown point, but not so large as to contain parameter values that cannot be rationalized with a well behaved P_1-density. To illustrate, consider again example 2.1; Y is a scalar, $\beta = E[Y] = p_D E_{P_1}[Y] + (1-p_D)E_{P_0}[Y]$, but for tractability suppose here that Kullback-Leibler is used to measure selection. Notice that since P_0 takes values on $[0,1]$, the Manski bounds for β are $\left[\frac{p_D}{2}, 1-\frac{p_D}{2}\right]$. Appendix A.1 shows that the dual problem has a solution whenever $b \in \left(\frac{p_D}{2}, 1-\frac{p_D}{2}\right)$, and that the solution satisfies strong duality. Thus for this example, B can be any set in the interior of the Manski bounds.

Assumption 2 is maintained throughout the remainder of the paper. Accordingly, the notation ν will be used for the value function of the dual problem as well.

4 Estimation and inference

The sample analogue of the dual problem provides an estimator of the value function, and suggests a simple plug-in estimator of the breakdown point. The asymptotic properties of these estimators are easier to study if the objective of the dual problem is expressed with a single unconditional expectation, which comes at the cost of additional notation.

First define the matrix $J(D) = \begin{bmatrix} -DI_{d_g} & 0 \\ 0 & (1-D)I_K \end{bmatrix}$ where I_{d_g} and I_K are identity matrices. Notice that $c(b) = \left(\frac{-p_D}{1-p_D}E[g(Y,X,b) \mid D=1]^\intercal \quad P(X=x_1 \mid D=0) \quad \ldots \quad P(X=x_K \mid D=0)\right)^\intercal = E\left[\frac{J(D)h(DY,X,b)}{(1-p_D)}\right]$ by the law of iterated expectations, and

$$\nu(b) = \sup_{\lambda \in \mathbb{R}^{d_g+K}} E\left[\frac{\lambda^\intercal J(D)h(DY,X,b)}{1-p_D} - \frac{Df^*\left(\lambda^\intercal h(DY,X,b)\right)}{p_D}\right] \tag{9}$$

Define

$$\varphi(D,DY,X,b,\lambda,p) = \frac{\lambda^\intercal J(D)h(DY,X,b)}{1-p} - \frac{D}{p}f^*(\lambda^\intercal h(DY,X,b)) \tag{10}$$

and observe that the dual problem is $\sup_{\lambda \in \mathbb{R}^{d_g+K}} E[\varphi(D,DY,X,b,\lambda,p_D)]$.

The estimator of the value function $\hat{\nu}_n : B \to \mathbb{R}$ is defined pointwise by

$$\hat{\nu}_n(b) = \sup_{\lambda \in \mathbb{R}^{d_g+K}} \frac{1}{n} \sum_{i=1}^{n} \varphi(D_i, D_i Y_i, X_i, b, \lambda, \hat{p}_{D,n}) \tag{11}$$

where $\hat{p}_{D,n} = \frac{1}{n}\sum_{i=1}^{n} D_i$ estimates p_D. Finally, $\hat{\delta}_n^{BP} = \inf_{b \in B \cap \mathbf{B}_0} \hat{\nu}_n(b)$ is the estimator of the breakdown point.

4.1 Consistency

The first consistency result notes that uniform consistency of $\hat{\nu}_n$ suffices to show consistency of $\hat{\delta}_n$. Theorem 4.2 then verifies uniform consistency of $\hat{\nu}_n$ in a wide class of GMM models, assuming the data possesses sufficient moments. These moments are also used in section 4.2 to derive the asymptotic distribution of the breakdown point and construct confidence intervals, but can be demanding in some cases. The consistency discussion concludes with Theorem 4.5, which shows that fewer moment conditions are needed to show consistency when the value function is convex.

Theorem 4.1 (Consistency of breakdown point)**.** *Suppose assumptions 1 and 2 hold, and* $\sup_{b \in B} |\hat{\nu}_n(b) - \nu(b)| \xrightarrow{p} 0$. *Then* $\hat{\delta}_n^{BP} \xrightarrow{p} \delta^{BP}$.

The proof, found in appendix F.1, is simply an application of the continuous mapping theorem.

4.1.1 Consistency in general GMM models

The following assumption suffices for consistency of $\hat{\nu}_n$, and is satisfied in many applications.

Assumption 3 (Consistency)**.** *Suppose that*

(i) B is compact

(ii) $g(z, b)$ is continuous in b for all z

(iii) For each $b \in B$, $\{h_j(y, x, b)\}_{j=1}^{d_g+K}$ are linearly independent in the sense that for any $\lambda \in \mathbb{R}^{d_g+K} \setminus \{0\}$,

$$P(\lambda^\intercal h(Y, X, b) \neq 0 \mid D = 1) > 0$$

(iv) For each $b \in B$, there exists closed convex $\bar{\Lambda}^b$ with $\lambda(b) \in int(\bar{\Lambda}^b)$ such that $\bar{\Lambda}^B := \{(b, \lambda) \ ; \ b \in$

$B, \lambda \in \bar{\Lambda}^b\}$ *is compact, and for some open set* $\mathcal{N} \subset \mathbb{R}$ *containing* p_D*,*

$$E\left[\sup_{p\in\mathcal{N}} \sup_{(b,\lambda)\in\bar{\Lambda}^B} |\varphi(D, DY, X, b, \lambda, p)|\right] < \infty,$$

$$E\left[\sup_{(b,\lambda)\in\bar{\Lambda}^B} \|\nabla_\lambda \varphi(D, DY, X, b, \lambda, p_D)\|\right] < \infty, \quad E\left[\sup_{(b,\lambda)\in\bar{\Lambda}^B} \|\nabla_\lambda^2 \varphi(D, DY, X, b, \lambda, p_D)\|\right] < \infty$$

The dual objective $E[\varphi(D, DY, X, b, \lambda, p_D)]$ is concave in λ for any distribution and any moment function g. Condition (iii) ensures this concavity strict, implying that $\lambda(b)$ solving the dual problem is unique. The moment conditions in (iv) are used to ensure $\lambda(b)$ is continuous in b, and used along with conditions (i) and (ii) to verify that the sample dual objective $\frac{1}{n}\sum_{i=1}^n \varphi(D_i, D_iY_i, X_i, b, \lambda, \hat{p}_{D,n})$ is uniformly consistent on $\bar{\Lambda}^B$.

Let $\hat{\lambda}_n(b) = \arg\max_\lambda \frac{1}{n}\sum_{i=1}^n \varphi(D_i, D_iY_i, X_i, b, \lambda, \hat{p}_{D,n})$. Although not a parameter of interest, uniform consistency of $\hat{\lambda}_n(b)$ is used to derive the asymptotic distribution of $\sqrt{n}(\hat{\nu}_n - \nu)$.

Theorem 4.2 (Consistency)**.** *Suppose assumptions 1, 2, and 3 hold. Then*

$$\sup_{b\in B} \|(\hat{\nu}_n(b), \hat{\lambda}_n(b)) - (\nu(b), \lambda(b))\| \xrightarrow{p} 0$$

The moment conditions in assumption 3 (iv) are similar to those in the econometrics literature utilizing divergences for estimation under misspecification, but can be demanding for some models and certain choices of divergence. Recall that $\varphi(D, DY, X, b, \lambda, p) = \frac{\lambda^\intercal J(D)h(DY,X,b)}{1-p} - \frac{D}{p} f^*\left(\lambda^\intercal h(DY, X, b)\right)$ and $\text{int}(\text{dom}(f^*)) = \text{int}(\{r \in \mathbb{R} \; ; \; f^*(r) < \infty\}) = (\ell^*, u^*)$. Assumption 3 (iv) implies that for all $(b, \lambda) \in \bar{\Lambda}^B$,

$$\ell^* < \lambda^\intercal h(DY, X, b) < u^*, \qquad\qquad P_1 - \text{a.s.}$$

Table 2 shows that for squared Hellinger, $u^* = 1/2$. Thus assumption 3 (iv) requires $\lambda^\intercal h(DY, X, b) < 1/2$ (P_1-a.s.) for all $(b, \lambda) \in \bar{\Lambda}^B$. This is easily violated if the support of (DY, X) and the moment functions g are unbounded. Assumption 3 (iv) is easier to satisfy if the divergence chosen to measure selection is Kullback-Leibler, which has $\ell^* = -\infty$ and $u^* = \infty$.[8]

[8] See Schennach (2007) and Broniatowski & Keziou (2012) remark 5.10 for additional discussion.

4.1.2 Consistency when the value function is convex

When $\hat{\nu}_n$ and ν are convex, uniform consistency of $\hat{\nu}_n$ comes without the demanding moment conditions of Assumption 3 (iv). For example, $\hat{\nu}_n$ and ν are convex when the parameter of interest is the expectation of a known function, or in linear models when only the outcome is missing.

Lemma 4.3 (Convex value function)**.** *Let B be convex, assumption 1 hold, and $g(y,x,b) = \tilde{g}(y,x) - b$. Then $\hat{\nu}_n$ and ν are convex on B. If assumption 2 holds as well, then ν is strictly convex on B.*

Lemma 4.4 (Convex value function, linear models)**.** *Let B be convex and assumption 1 hold. Suppose the sample is $\{D_i, D_iY_i, X_{i1}, X_{i2}\}_{i=1}^n$, where $Y_i \in \mathbb{R}$, $X_{i1} \in \mathbb{R}^{d_{x1}}$, and $X_{i2} \in \mathbb{R}^{d_{x2}}$ with $d_{x2} \geq d_{x1}$. Consider an instrumental variables model:*

$$Y_i = X_{i1}^\intercal \beta + \varepsilon, \qquad\qquad E\left[X_{i2}\varepsilon\right] = 0$$

Then $\hat{\nu}_n$ and ν are convex on B.

Ordinary least squares can be recovered as the special case of lemma 4.4 when $X_2 = X_1$. Simulations suggest that OLS more generally produces convex $\nu(b)$, as discussed in appendix A.2.

Remark 4.1. It is worth noting that the setting of lemma 4.4 encompases the popular local average treatment effect framework of Imbens & Angrist (1994) (example 3). Specifically, suppose $\tilde{Z}_i \in \{0,1\}$ indicates assignment to treatment and $T_i \in \{0,1\}$ indicates receipt of treatment. These are typically observed for every individual i, and thus $X_{i1} = \begin{pmatrix}1, & T_i\end{pmatrix}^\intercal$ and $X_{i2} = \begin{pmatrix}1 & \tilde{Z}_i\end{pmatrix}^\intercal$ are finitely supported and always observed. Most trials suffer from some level of attrition, resulting in occasionally missing outcomes of interest Y_i.

Convexity and pointwise consistency of $\hat{\nu}_n$ on an open convex set implies uniform consistency on compact subsets of that open set. (See Rockafellar (1970) theorem 10.8, and Andersen & Gill (1982) theorem II.1). This bypasses the need for the moment conditions in Assumption 3 (iv).

Theorem 4.5 (Consistency with convex value function)**.** *Suppose that assumptions 1 and 2 hold, and*

(i) B is a convex, compact subset of int($\boldsymbol{B}$)

(ii) For each $b \in B$, $\{h_j(y,x,b)\}_{j=1}^{d_g+K}$ are linearly independent in the sense that for any $\lambda \in \mathbb{R}^{d_g+K} \setminus \{0\}$,

$$P(\lambda^\intercal h(Y,X,b) \neq 0 \mid D = 1) > 0$$

(iii) $\hat{\nu}_n(\cdot)$ and $\nu(\cdot)$ are convex

Then

$$\sup_{b \in B} |\hat{\nu}_n(b) - \nu(b)| \xrightarrow{p} 0$$

It is worth noting that while these weaker conditions deliver consistency of the breakdown point, the theorem does not imply uniform consistency of $\hat{\lambda}_n(b) = \arg\max_\lambda \hat{\nu}_n(b, \lambda)$. Uniform consistency of $\hat{\lambda}_n(b)$ is used in theorem 4.6 to establish the asymptotic distribution of $\sqrt{n}(\hat{\nu}_n - \nu)$.

4.2 Inference

A large breakdown point implies the incomplete distribution P_0 would have to differ greatly from P_1 to rationalize the null hypothesis. If δ^{BD} is larger than the plausible amount of selection in the setting, the null hypothesis is similarly implausible. Skeptical readers following this argument may worry the point estimate $\hat{\delta}_n^{BP}$ is larger than δ^{BP} due to sample noise – but the force of the argument is only strengthened if $\hat{\delta}_n^{BP}$ falls below δ^{BP}. To address these concerns, researchers should report lower confidence intervals along with point estimates of the breakdown point.

This section constructs consistent lower confidence interval. Such an interval is given by $\widehat{CI}_L$ such that

$$\lim_{n \to \infty} P\left(\widehat{CI}_L \leq \delta^{BD}\right) \geq 1 - \alpha$$

The confidence interval constructed below is based on the asymptotic distribution of $\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})$. Recall that $\delta^{BP} = \inf_{b \in B \cap \mathbf{B}_0} \nu(b)$, and $\hat{\delta}_n^{BP} = \inf_{b \in B \cap \mathbf{B}_0} \hat{\nu}_n(b)$. The plug-in structure suggests that the functional delta method can be used to derive the asyptotic distribution of $\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})$ from the asymptotic distribution of $\sqrt{n}(\hat{\nu}_n - \nu)$.

4.2.1 Asymptotic distribution of the value function

The asymptotic distribution of $\sqrt{n}(\hat{\nu}_n - \nu)$ follows from a uniform linearization. To gain intuition for this result, consider estimating $\sqrt{n}(\hat{\nu}_n(b) - \nu(b))$ for a particular b. The estimators $\hat{\nu}_n(b)$, $\hat{\lambda}_n(b)$, and $\hat{p}_{D,n}$ can be viewed as a single GMM estimator of $(\nu(b), \lambda(b), p_D)$:

$$0 = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{p}_{D,n}) - \hat{\nu}_n(b) \\ \nabla_\lambda \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{p}_{D,n}) \\ D_i - \hat{p}_{D,n} \end{pmatrix}$$

Well known results imply $(\hat{\nu}_n(b), \hat{\lambda}_n(b), \hat{p}_{D,n})$ is asymptotically linear (see, e.g., Newey & McFadden (1994)). Theorem 4.6 below ensures this linearization holds uniformly over B, from which the asymptotic distribution of $\sqrt{n}(\hat{\nu}_n - \nu)$ follows.

For notational convenience, group the estimands:

$$\theta_0 : B \to \mathbb{R}^{d_g+K+2}, \qquad\qquad \theta_0(b) = \begin{pmatrix} \nu(b) & \lambda(b)^\intercal & p_D \end{pmatrix}^\intercal$$

and the corresponding estimators $\hat{\theta}_n(b) = \begin{pmatrix} \hat{\nu}_n(b) & \hat{\lambda}_n(b)^\intercal & \hat{p}_{D,n} \end{pmatrix}^\intercal$. Group the moment functions used in estimating $\theta_0(b)$ together in ϕ:

$$\phi(D, DY, X, b, \theta) = \phi(D, DY, X, b, v, \lambda, p) = \begin{pmatrix} \varphi(D, DY, X, b, \lambda, p) - v \\ \nabla_\lambda \varphi(D, DY, X, b, \lambda, p) \\ D - p \end{pmatrix}, \tag{12}$$

and define the corresponding Jacobian term $\Phi(b) = E\left[\nabla_\theta \phi(D, DY, X, b, \theta_0)\right]$. Lastly define the parameter space: for $\mathcal{V} > 0$ and $[\underline{p}, \overline{p}] \subset (0, 1)$, let

$$\Theta^b = \left\{\theta = (v, \lambda, p) \; ; \; v \in [0, \mathcal{V}], \lambda \in \bar{\Lambda}^b, p \in [\underline{p}, \overline{p}]\right\} \tag{13}$$

and set $\Theta^B = \left\{(b, \theta) \; ; \; b \in B, \theta \in \Theta^b\right\}$.

Assumption 4 (Inference)**.** *Suppose that*

(i) $\boldsymbol{B}_0$ *is closed*

(ii) B *is convex*

(iii) $g(z, b)$ *is continuously differentiable with respect to* b

(iv) $\hat{\theta}_n(b) = (\hat{\nu}_n(b), \hat{\lambda}_n(b), \hat{p}_{D,n}) \in \Theta^b$ *for each* b

(v) There exists $F(d, dy, x)$ *such that*

$$\sup_{b \in B} \sup_{\theta \in \Theta^b} \|\nabla_{(b,\theta)} \phi(d, dy, x, b, \theta)\| \leq F(d, dy, x)$$

and $E[F(D, DY, X)^2] < \infty$.

Condition (i) ensures $B \cap \mathbf{B}_0$ is compact, which is used when applying the functional delta method. Conditions (ii) through (v) ensure certain classes of functions are Donsker.[9] Conditions (iv) and (v) are also used to ensure Jacobian terms are invertible and that the sample analogues of those Jacobian term inverses are uniformly consistent.

[9] Convexity of B ensures convexity of Θ^B, which is used in lemma F.8 to apply the mean value theorem when verifying a particular class of functions is Donsker.

The statement of the following theorem requires some additional notation. $\sqrt{n}(\hat{\theta}_n - \theta_0)$ can be viewed as vector-valued processes on B with marginals in $\mathbb{R}^{d_g+K+2}$, or as scalar valued processes on $B \times \mathcal{I}$ where $\mathcal{I} = \{1, \ldots, d_g + K + 2\}$ indexes the coordinates of $\mathbb{R}^{d_g+K+2}$. Let $\ell^\infty(B \times \mathcal{I})$ be the set of real-valued bounded functions with domain $B \times \mathcal{I}$ equipped with the supremum norm. Lastly, $\xrightarrow{L}$ denotes convergence in law.

Theorem 4.6 (Asymptotic distribution of the value function)**.** *Suppose assumptions 1, 2, 3, and 4 hold. Then as a process in $\ell^\infty(B \times \mathcal{I})$,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathbb{G}$$

where $\mathbb{G}$ is a mean zero tight Gaussian process, with covariance function

$$\begin{aligned} &Cov(\mathbb{G}(b_1, i_1), \mathbb{G}(b_2, i_2)) \\ &\quad = E\left[\left(\Phi(b_1)^{-1}\right)^{(i_1)} \phi\left(D, DY, X, b_1, \theta(b_1)\right) \left[\left(\Phi(b_2)^{-1}\right)^{(i_2)} \phi\left(D, DY, X, b_2, \theta(b_2)\right)\right]\right] \end{aligned}$$

where $\left(\Phi(b)^{-1}\right)^{(i)}$ is the i-th row of the matrix $\Phi(b)^{-1} = E\left[\nabla_\theta \phi(D, DY, X, b, \theta_0)\right]^{-1}$.

Marginalizing to the first coordinate implies the asymptotic distribution of the value function. Specifically, let $\mathbb{G}_\nu$ be defined pointwise as $\mathbb{G}_\nu(b) = \mathbb{G}(b, 1)$. Theorem 4.6 and the continuous mapping theorem implies $\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu$.

4.2.2 Asymptotic distribution of the breakdown point

The asymptotic distribution of $\sqrt{n}(\hat{\delta}^{BP} - \delta^{BP})$ follows from Theorem 4.6 and the functional delta method for Hadamard directionally differentiable functions (Fang & Santos (2019)).

Theorem 4.7 (Asymptotic distribution of the breakdown point)**.** *Suppose assumptions 1, 2, 3, and 4 hold. Then*

$$\sqrt{n}(\hat{\delta}^{BP}_n - \delta^{BP}) \xrightarrow{L} \inf_{b \in \boldsymbol{m}(\nu)} \mathbb{G}_\nu(b)$$

where $\boldsymbol{m}(\nu) = \arg\min_{b \in B \cap \boldsymbol{B}_0} \nu(b)$.

Note that if $\mathbf{m}(\nu) = \arg\min_{b \in B \cap \mathbf{B}_0} \nu(b)$ is the singleton $\{b^i\}$, $\inf_{b \in \mathbf{m}(\nu)} \mathbb{G}_\nu(b)$ simplifies to $\mathbb{G}_\nu(b^i)$. In this case, $\sqrt{n}(\hat{\delta}^{BP}_n - \delta^{BP})$ is asymptotically normal. This appears quite plausible in many applications, and is guaranteed to hold when $\mathbf{B}_0$ is convex and ν is strictly convex - as in the example studied in lemma 4.3.

4.2.3 Bootstrap

Consistency of the lower confidence interval will follow from consistent estimation of the distribution of $\inf_{b\in\mathbf{m}(\nu)}\mathbb{G}_\nu(b)$. This can be accomplished with the tools studied in Fang & Santos (2019), by first consistently estimating the distribution $\mathbb{G}_\nu$.

The "score bootstrap" is a computationally attractive method to estimate the distribution of $\mathbb{G}$. This procedure works by perturbing an estimate of the influence function with random weights $\{W_i\}_{i=1}^n$. The distribution of these weights is chosen so as to maintain the first and second moments of the limiting distribution.[10]

Assumption 5 (Score bootstrap)**.** *$\{W_i\}_{i=1}^n$ are i.i.d. scalars, independent of $\{D_i, D_iY_i, X_i\}_{i=1}^n$, satisfying (i) $E[W]=0$, (ii) $E[W^2]=1$, and (iii) $E[|W|^{2+a}]<\infty$ for some $a>0$.*

Theorem 4.8 (Score bootstrap consistency)**.** *Suppose assumptions 1, 2, 3, and 4 hold, and $\{W_i\}_{i=1}^n$ satisfies assumption 5. Let $\hat{\Phi}_n(b)=\frac{1}{n}\sum_{i=1}^n\nabla_\theta\phi(D_i,D_iY_i,X_i,b,\hat{\theta}_n(b))$, and $\hat{G}_n^*:B\to\mathbb{R}^{d_g+K+2}$ be defined pointwise as*

$$\hat{G}_n^*(b)=\hat{\Phi}_n(b)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(D_i,D_iY_i,X_i,b,\hat{\theta}_n(b)) \tag{14}$$

Further define $\hat{\theta}_n^*(b)=\frac{1}{\sqrt{n}}\hat{G}_n^*(b)+\hat{\theta}_n(b)$. Then conditional on $\{D_i,D_iY_i,X_i\}_{i=1}^n$,

$$\sqrt{n}(\hat{\theta}_n^*-\hat{\theta}_n)\xrightarrow{L}\mathbb{G}$$

in outer probability.

Theorem 4.8 directly implies the score bootstrap is consistent for $\mathbb{G}_\nu$. Specifically, define $\hat{\nu}_n^*:B\to\mathbb{R}$ pointwise with $\hat{\nu}_n^*(b)=\frac{1}{\sqrt{n}}\hat{G}_n^*(b,1)+\hat{\theta}_n(b,1)$. Then conditional on $\{D_i,D_iY_i,X_i\}_{i=1}^n$, $\sqrt{n}(\hat{\nu}_n^*-\hat{\nu}_n)\xrightarrow{L}\mathbb{G}_\nu$ in outer probability.

4.2.4 Lower confidence intervals

A consistent, $1-\alpha$ lower confidence interval is characterized by $\widehat{CI}_L$ such that

$$\lim_{n\to\infty}P\left(\widehat{CI}_L\leq\delta^{BD}\right)\geq 1-\alpha.$$

[10] To the author's knowledge, the score bootstrap procedure was proposed by Lewbel (1995) and Hansen (1996). The result here shows consistency of the score bootstrap for a stochastic process, and is similar to that shown in Kaido & Santos (2014).

Consistent confidence intervals can be constructed by $\widehat{CI}_L = \hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}}\hat{c}_{1-\alpha,n}$ where $\hat{c}_{1-\alpha,n}$ is a suitable estimator of $c_{1-\alpha} = \inf\left\{c\ ;\ P\left(\inf_{b\in\mathbf{m}(\nu)}\mathbb{G}_\nu(b) \leq c\right) \geq 1-\alpha\right\}$. The bootstrap is used to construct such estimators.

As mentioned above, in most cases it is plausible that $\mathbf{m}(\nu) = \arg\min_{b\in B\cap\mathbf{B}_0}\nu(b)$ is a singleton. In this case, Fang & Santos (2019) theorem 3.1 implies the "plug-in bootstrap" works to produce a consistent confidence interval, with quantile $\hat{c}_{1-\alpha,n}^{Plug}$.[11] Although $\hat{c}_{1-\alpha,n}^{Plug}$ can be computed numerically through simulation, doing so can be computationally intensive as $\inf_{b\in B\cap\mathbf{B}_0}\hat{\nu}_n^*(b)$ must be solved for each bootstrap sample. The following result provides a computationally attractive alternative.

Theorem 4.9 (Confidence interval consistency)**.** *Suppose assumptions 1, 2, 3, and 4 hold, and* $\{W_i\}_{i=1}^n$ *satisfies assumption 5. Let* $\hat{G}_n^*$ *be as defined in theorem 4.8, and* $\hat{\nu}_n^* : B \to \mathbb{R}$ *defined pointwise by* $\hat{\nu}_n(b) = \frac{1}{\sqrt{n}}\hat{G}_n(b,1) + \hat{\theta}_n(b,1)$. *Let* $\hat{b}_n^i$ *solve* $\min_{b\in B\cap\boldsymbol{B}_0}\hat{\nu}_n(b)$, *and*

$$\hat{c}_{1-\alpha,n}^{Simple} = \inf\left\{c\ ;\ P\left(\sqrt{n}(\hat{\nu}_n^*(\hat{b}_n^i) - \hat{\nu}_n(\hat{b}_n^i)) \leq c \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right) \geq 1-\alpha\right\}$$

If $\boldsymbol{m}(\nu) = \arg\min_{b\in B\cap\boldsymbol{B}_0}\nu(b)$ *is the singleton* $\{b^i\}$, *then*

$$\lim_{n\to\infty} P\left(\hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}}\hat{c}_{1-\alpha,n}^{Simple} \leq \delta^{BP}\right) = 1-\alpha$$

Computing $\hat{c}_{1-\alpha,n}^{Simple}$ through simulation is straightforward and fast. The following steps can be used to compute the estimator and construct a consistent lower confidence interval.

1. Compute $\hat{b}_n^i = \arg\min_{b\in B\cap\mathbf{B}_0}\hat{\nu}_n(b)$, where $\hat{\nu}_n(b)$ is defined in (11).

2. Generate N bootstrap samples $\{W_i\}_{i=1}^n$ from a distribution satisfying assumption 5, and compute $\hat{G}_n^*(\hat{b}_n^i, 1)$ for each of the N bootstrap samples using (14).

3. Let $\hat{c}_{1-\alpha,n}^{Simple}$ be the $1-\alpha$ quantile of $\{\hat{G}_{n,k}^*(\hat{b}_n^i, 1)\}_{k=1}^N$.

The point estimate of the breakdown point is then given by $\hat{\delta}_n^{BP} = \hat{\nu}_n(\hat{b}_n^i)$, and the $1-\alpha$ lower confidence interval characterized by $\widehat{CI}_L = \hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}}\hat{c}_{1-\alpha,n}^{Simple}$.

Remark 4.2. Theorem 4.9 assumes $\arg\min_{B\cap\mathbf{B}_0}\nu(b)$ is unique. This is quite plausible for most GMM models, and guaranteed in the example studied in lemma 4.3. However, if $\arg\min_{B\cap\mathbf{B}_0}\nu(b)$ is not unique it appears plausible that confidence intervals based on $\hat{c}_{1-\alpha,n}^{Simple}$ will be asymptotically conservative, in the sense that $\liminf_{n\to\infty} P\left(\hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}}\hat{c}_{1-\alpha,n}^{Simple} \leq \delta^{BP}\right) \geq 1-\alpha$. The idea behind this conjecture is that when $\mathbf{m}(\nu)$ is not unique, the random variable $\sqrt{n}\left(\hat{\nu}_n^*(\hat{b}_n^i) - \hat{\nu}_n(\hat{b}_n^i)\right)$ is

[11] See lemma F.11 in appendix F.2.3 for a formal definition.

(asymptotically) larger than $\inf_{b \in \mathbf{m}(\nu)} \mathbb{G}_{\nu}(b)$. This would imply $\hat{c}^{Simple}_{1-\alpha,n}$ is "too big," and $\hat{\delta}^{BP}_{n} - \frac{1}{\sqrt{n}} \hat{c}^{Simple}_{1-\alpha,n}$ is lower than it needs to be.

5 Simulations

To demonstrate the finite sample properties of the proposed estimators, this section presents simulation results on a variety of different data generating processes.

5.1 Simple mean

Recall example 2.1. The parameter of interest is the mean of a scalar random variable Y, $\beta = E[Y] = p_D E_{P_1}[Y] + (1 - p_D) E_{P_0}[Y]$, and the sample is $\{D_i, D_i Y_i\}_{i=1}^{n}$. The observed distribution of $Y \mid D = 1$ is the uniform distribution on $[0, 1]$. The probability of observing Y is $p_D = P(D = 1) = 0.7$. To support the claim $H_1 : \beta > 0.4$, let $H_0 : \beta \leq 0.4$. Selection is measured using squared Hellinger, $d_f(Q \| P_1) = H^2(Q, P_1)$. Recall that the true breakdown point, δ^{BP}, of this example is just over 0.2.

The following table summarizes simulations 250 simulations for several different sample size.[12]

Table 3: Simulations, Squared Hellinger, Uniform, Mean

n	RMSE	Emp. Bias	Emp. CI Coverage	Ave. CI Length
1000	0.060	0.008	98.4	0.091
2000	0.040	0.005	97.6	0.063
3000	0.032	0.001	96.8	0.051
5000	0.024	0.003	96.4	0.040

The simulations show little bias. Coverage is slightly above the targeted 95% significance level in smaller samples.

5.2 Linear Models

Linear models are the among the most common tools used by empirical researchers. This subsection uses simulations to investigate linear regression with exogenous regressors.[13]

[12] Here Emp. Bias = $\frac{1}{250} \sum_{s=1}^{250} (\hat{\delta}^{BP}_{n,s} - \delta^{BP})$ and Ave. CI Length = $\frac{1}{250} \sum_{s=1}^{250} (\hat{\delta}^{BP}_{n,s} - \widehat{CI}_{L,s})$.

[13] Recall that lemma 4.4 proves that when the outcome of a regression is the only missing variable, $\nu(\cdot)$ is convex. Appendix A.2 shows simulation evidence that the $\nu(\cdot)$ of the following DGP is convex.

Consider the following model:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 Y_2 + \beta_3 X_2 + \varepsilon = W^\intercal \beta + \varepsilon, \tag{15}$$

where $W = \begin{pmatrix} 1 & X_1 & Y_2 & X_2 \end{pmatrix}^\intercal$ are the exogenous regressors: $E[W\varepsilon] = 0$. Here Y_1 is a continuous outcome variable, $X_1 = \{0, 1\}$ is the regressor of interest, Y_2 is a continuously distributed control, and $X_2 \in \{0, 1, 2\}$ is a discrete control. The conclusion whose robustness is to be investigated is that the coefficient on X_1 is positive:

$$H_0 \ : \ \beta_1 \leq 0, \qquad\qquad H_1 \ : \ \beta_1 > 0$$

The researcher observes the sample $\{D_i, D_i Y_{i1}, D_i Y_{i2}, X_{i1}, X_{i2}\}_{i=1}^n$ and once again uses squared Hellinger to measure selection.

The data generating process specification takes inspiration from mincerian wage equations. For worker i, let Y_{i1} be i's log-income, X_{i1} an indicator for i being a college graduate, Y_{i2} i's work experience, and X_{i2} the number of parents with college degrees (0, 1, or 2). Specifically, let X_2 be multinomial,[14] $X_1 \sim \text{Binomial}\left(\frac{X_2+1}{4}\right)$, and $Y_2 \sim \text{Beta}(3 - X_1, 3)$. Let $\tilde{\varepsilon} \sim U[-1, 1]$ (independent of all other variables), $\varepsilon = (X_1 + 1)\tilde{\varepsilon}$. $\beta_0 = \beta_1 = \beta_2 = 1$, and $\beta_3 = 0.5$. Finally, Y_1 generated according to (15). For the missing data process, let $D = \{\varepsilon X_1 + 10X_1 + 10(X_2 - 1) > \eta\}$, where $\eta \sim N(\mu_\eta, \sigma_\eta^2)$. In the simulations, $\mu_\eta = -9$ and $\sigma_\eta = 15$, which implies $P(D = 1) \approx 0.7$. Notice the support of (Y_1, Y_2, X_1, X_2) is compact, ensuring the moment conditions in assumptions 3 and 4 are satisfied.

Unlike in the uniform mean example, the population value of the breakdown point is difficult to compute. For these simulations, the population value is approximated as the point estimate obtained from a very large sample. Specifically, the point estimate from 1 million observations is about 0.2068.

The following table summarizes 250 simulations for several different sample sizes.

[14] $P(X_2 = 0) = 0.5$, $P(X_2 = 1) = 0.3$, and $P(X_2 = 2) = 0.2$

Table 4: Simulations, Squared Hellinger, OLS

n	RMSE	Emp. Bias	Emp. CI Coverage	Ave. CI Length
1000	0.043	0.009	100.0	0.078
2000	0.033	0.005	98.0	0.052
3000	0.026	0.007	98.0	0.043
5000	0.017	0.002	98.0	0.032

The simulations again show little bias, but confidence interval coverage is noticably above the targeted 95% significance level and does not appear to be converging. This may be due to the conjecture that inference with this procedure will be conservative; see remark 4.2. If $p_D P_1 + (1 - p_D)Q$ is such that the design matrix is not invertible, there will be several parameters values b solving the moment condition.

6 Conclusion

This paper proposes breakdown point analysis as a tractable approach to assessing the sensitivity of a researcher's conclusion to the common MCAR assumption. When defined with squared Hellinger, the breakdown point δ^{BP} has a natural interpretation: if the result were false, the variables under study (Z) would have to predict an observation being selected into the sample (D) at least well enough that $H^2(P_0, P_1) = 1 - E[\sqrt{\text{Var}(D \mid Z)}]/\sqrt{\text{Var}(D)} \geq \delta^{BP}$. Simple estimators and bootstrap procedures are proposed and shown consistent, allowing for lower confidence intervals to be constructed. Researchers working with incomplete datasets should report the breakdown point estimate and lower confidence interval along with standard results, making transparent to their audience how robust the conclusion is to relaxing the MCAR assumption.

References

Aliprantis, C. D., & Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, (pp. 1100–1120).

Bhatia, R. (1997). *Matrix Analysis*. New York, NY: Springer New York.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., & Ziliak, J. P. (2019). Trouble in the tails? what we know about earnings nonresponse 30 years after lillard, smith, and welch. *Journal of Political Economy*, *127*(5), 2143–2185.

Borwein, J. M., & Lewis, A. S. (1991). Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, *29*(2), 325–338.

Borwein, J. M., & Lewis, A. S. (1993). Partially-finite programming in l_1 and the existence of maximum entropy estimates. *SIAM Journal on Optimization*, *3*(2), 248–267.

Brame, R., Turner, M. G., Paternoster, R., & Bushway, S. D. (2012). Cumulative prevalence of arrest from ages 8 to 23 in a national sample. *Pediatrics*, *129*(1), 21–27.

Broniatowski, M., & Keziou, A. (2006). Minimization of φ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, *43*(4), 403–442.

Broniatowski, M., & Keziou, A. (2012). Divergences and duality for estimation and test under moment condition models. *Journal of Statistical Planning and Inference*, *142*(9), 2554–2573.

Coleman, R. (2012). *Calculus on normed vector spaces*. Springer Science & Business Media.

Csiszár, I., Gamgoa, F., & Gassiat, E. (1999). Mem pixel correlated solutions for generalized moment and interpolation problems. *IEEE Transactions on Information Theory*, *45*(7), 2253–2270.

Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, *70*(1), 33–58.

Diegert, P., Masten, M. A., & Poirier, A. (2022). Assessing omitted variable bias when the controls are endogenous. *arXiv preprint arXiv:2206.02303*.

European Parliament and Council of the European Union (2016). Regulation (eu) 2016/679. URL http://data.europa.eu/eli/reg/2016/679/2016-05-04

Fang, Z., & Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, *86*(1), 377–412.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica: Journal of the econometric society*, (pp. 413–430).

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, (pp. 153–161).

Horowitz, J. L., & Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, (pp. 281–302).

Horowitz, J. L., & Manski, C. F. (2006). Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics*, *132*(2), 445–459.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica: journal of the Econometric Society*, (pp. 467–475).

Isaac, M. (2022). Meta spent $10 billion on the metaverse in 2021, dragging down profit. *The New York Times*.
URL https://www.nytimes.com/2022/02/02/technology/meta-facebook-earnings-metaverse.html

Kaido, H., & Santos, A. (2014). Asymptotically efficient estimation of models defined by convex moment inequalities. *Econometrica*, *82*(1), 387–413.

Kline, P., & Santos, A. (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the us wage structure. *Quantitative Economics*, *4*(2), 231–267.

Lewbel, A. (1995). Consistent nonparametric hypothesis tests with an application to slutsky symmetry. *Journal of Econometrics*, *67*(2), 379–401.

Manski, C. F. (2005). Partial identification with missing data: concepts and findings. *International Journal of Approximate Reasoning*, *39*(2-3), 151–165.

Manski, C. F. (2013). Response to the review of 'public policy in an uncertain world'.

Masten, M. A., & Poirier, A. (2020). Inference on breakdown frontiers. *Quantitative Economics*, *11*(1), 41–111.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Rockafellar, R. T. (1970). *Convex analysis*, vol. 18. Princeton university press.

Satariano, A. (2019). Google is fined $57 million under europe's data privacy law. *The New York Times*.
URL https://www.nytimes.com/2019/01/21/technology/google-europe-gdpr-fine.html

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, *35*(2), 634–672.

van der Vaart, A., & Wellner, J. A. (1997). *Weak convergence and empirical processes with applications to statistics*. London: Royal Statistical Society, 1988-.

van der Vaart, A. W. (2007). *Asymptotic statistics*. Cambridge university press.

Zeidler, E. (1986). Nonlinear functional analysis vol. 1: Fixed-point theorems.

Appendices

Contents

A Examples

A.1 Expectation

This simple example is useful primarily to illustrate the ideas in a concrete setting.

Suppose the parameter of interest is $\beta = E[Y] \in \mathbb{R}$, and the sample is $\{D_i, D_iY_i\}_{i=1}^n$. The conclusion to be supported is that $\beta > \bar{b}$, motivating the null and alternative hypotheses

$$H_0 \ : \ \beta \leq \bar{b}, \qquad\qquad H_1 \ : \ \beta > \bar{b}$$

The model is characterized by $g(y, b) = y - b$. For the dual problem, set $h(y, b) = \begin{pmatrix} y - b & 1 \end{pmatrix}^\intercal$. The dual problem is

$$\sup_{\lambda \in \mathbb{R}^2} \nu(b, \lambda) = \sup_{\lambda \in \mathbb{R}^2} \lambda^\intercal c(b) - E_{P_1}[f^*(\lambda^\intercal h(Y, b))] \tag{16}$$

where $c(b) = \begin{pmatrix} \frac{-p_D}{1-p_D}(E_{P_1}[Y] - b) & 1 \end{pmatrix}^\intercal$.

A.1.1 Dual solution when d_f is Kullback-Leibler and P_1 is $\mathcal{U}[0, 1]$

Suppose that P_1, the distribution of $Y \mid D = 1$, is $\mathcal{U}[0, 1]$. Let $\mu_1 = E[Y \mid D = 1] = 1/2$. Note that, since the support of P_0 is contained within $[0, 1]$ as well, we have $\beta = E[Y] \in [p_D\mu_1, p_D\mu_1 + (1 - p_D)]$. The endpoints are only attained if P_0 concentrates degenerately at 0 or 1 respectively, distributions which violate $P_0 \ll P_1$.

For tractability, let the measure of selection be Kullback-Leibler. For this divergence we let $f(t) = t\log(t) - t + 1$, which has convex conjugate $f^*(r) = \exp(r) - 1$. The dual problem has first order condition

$$\begin{aligned} 0 &= c(b) - E_{P_1}\left[(f^*)'(\lambda^\intercal h(Y, b))h(Y, b)\right] \\ &= \begin{pmatrix} \frac{-p_D}{1-p_D}\left(\frac{1}{2} - b\right) \\ 1 \end{pmatrix} - E\left[\exp\left(\lambda_1(Y - b) + \lambda_2\right)\begin{pmatrix} (Y - b) \\ 1 \end{pmatrix}\right] \end{aligned}$$

From the second equation we have

$$\lambda_2 = -\log\left(E[\exp(\lambda_1(Y - b))]\right) \tag{17}$$

Suppose $b = \frac{1}{2}$. Then the first equation requires

$$0 = E\left[\exp(\lambda_1(Y - b) + \lambda_2)\left(Y - \frac{1}{2}\right)\right] \tag{18}$$

Notice that if $\lambda_1 = 0$, then (17) implies $\lambda_2 = 0$, and (18) holds.

Now suppose $b \neq 1/2$. Consider the dual objective, and notice that

$$E_{P_1}[f^*(\lambda^\intercal h(Y,b))] = \int_0^1 \exp(\lambda^\intercal h(y,b)) - 1dy$$

Since $b \neq 1/2$, it follows that $\frac{-p_D}{1-p_D}(1/2 - b) \neq 0$ and so $\lambda_1 \neq 0$. Thus the integral above can be solved with u-substition, setting $u = \lambda_1(y-b) + \lambda_2$:

$$\begin{aligned} E_{P_1}[f^*(\lambda^\intercal h(Y,b))] &= \int_0^1 \exp(\lambda^\intercal h(y,b)) - 1dy = \frac{1}{\lambda_1}\int_{\lambda_1(-b)+\lambda_2}^{\lambda_1(1-b)+\lambda_2} \exp(u)du - 1 \\ &= \frac{\exp(\lambda^\intercal \mathbf{b}_1) - \exp(\lambda^\intercal \mathbf{b}_0)}{\lambda^\intercal e_1} - 1 \end{aligned}$$

where $\mathbf{b}_1 = \begin{pmatrix}1-b & 1\end{pmatrix}^\intercal$, $\mathbf{b}_0 = \begin{pmatrix}-b & 1\end{pmatrix}^\intercal$, and $e_1 = \begin{pmatrix}1 & 0\end{pmatrix}^\intercal$. Thus (16) becomes

$$\sup_{\lambda\in\mathbb{R}^2} \lambda^\intercal \begin{pmatrix}\frac{-p_D}{1-p_D}(1/2-b) \\ 1\end{pmatrix} - \frac{\exp(\lambda^\intercal \mathbf{b}_1) - \exp(\lambda^\intercal \mathbf{b}_0)}{\lambda^\intercal e_1} + 1$$

from which we can compute the first order conditions

$$0 = \begin{pmatrix}\frac{-p_D}{1-p_D}(1/2-b) \\ 1\end{pmatrix} - \frac{\exp(\lambda^\intercal \mathbf{b}_1)\mathbf{b}_1 - \exp(\lambda^\intercal \mathbf{b}_0)\mathbf{b}_0}{\lambda^\intercal e_1} + \frac{\exp(\lambda^\intercal \mathbf{b}_1) - \exp(\lambda^\intercal \mathbf{b}_0)}{(\lambda^\intercal e_1)^2} e_1$$

Once again, the second equation can be solved for λ_2. The following form will be more useful:

$$\begin{aligned} 0 &= 1 - \frac{\exp(\lambda_1(1-b)+\lambda_2) - \exp(\lambda_1(-b)+\lambda_2)}{\lambda_1} \\ \implies \frac{\lambda_1}{\exp(\lambda_2)} &= \exp(\lambda_1(1-b)) - \exp(\lambda_1(-b)) \end{aligned} \tag{19}$$

The first equation is

$$\begin{aligned} \frac{-p_D}{1-p_D}\left(\frac{1}{2} - b\right) &= \frac{\exp(\lambda_1(1-b)+\lambda_2)(1-b) - \exp(\lambda_1(-b)+\lambda_2)(-b)}{\lambda_1} \\ &\qquad - \frac{\exp(\lambda_1(1-b)+\lambda_2) - \exp(\lambda_1(-b)+\lambda_2)}{\lambda_1^2} \\ &= \frac{\exp(\lambda_2)}{\lambda_1}\Bigg[\exp(\lambda_1(1-b)) - b[\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))] \\ &\qquad - \frac{\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))}{\lambda_1}\Bigg] \\ &= \frac{\exp(\lambda_1(1-b))}{\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))} - b - \frac{1}{\lambda_1} = \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - b - \frac{1}{\lambda_1} \end{aligned}$$

where the second to last equality uses (19) above. Rearranging gives

$$\frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = \frac{-p_D(1/2-b) + (1-p_D)b}{1-p_D} = \frac{2b - p_D}{2(1-p_D)}$$

Now notice that $\frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1}$ is well defined and continuous whenever $\lambda_1 \neq 0$, takes values between 0 and 1, with limits

$$\lim_{\lambda_1 \to \infty} \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = 1, \qquad \lim_{\lambda_1 \to -\infty} \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = 0$$

Repeated applications of l'Hôpital's rule shows that

$$\lim_{\lambda_1 \to 0} \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = \lim_{\lambda_1 \to 0} \frac{\lambda_1 \exp(\lambda_1) - \exp(\lambda_1) + 1}{\lambda_1(\exp(\lambda_1) - 1)} = \frac{1}{2}$$

Therefore there exists a solution whenever $\frac{2b-p_D}{2(1-p_D)} \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right)$. Given this solution, (19) can be rearranged to obtain

$$\lambda_2 = \log\left(\frac{\lambda_1}{\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))}\right)$$

Now notice that

$$\frac{2b-p_D}{2(1-p_D)} > 0 \implies b > \frac{p_D}{2},$$
$$\frac{2b-p_D}{2(1-p_D)} < 1 \implies b < 1 - \frac{p_D}{2}$$

and recall that $b = 1/2$ implies $\lambda_1 = \lambda_2 = 0$ solves the dual problem. Therefore the dual problem has a solution whenever $b \in \left(\frac{p_D}{2}, 1 - \frac{p_D}{2}\right)$.

P_1 has compact support, and $f^*(\lambda_1(y-b) + \lambda_2) = \exp(\lambda_1(y-b) + \lambda_2) - 1$ is continuous in y for any (λ_1, λ_2). Thus the extreme value theorem implies the solution is in the interior of $\left\{\lambda \in \mathbb{R}^2 \; ; \; E[|f^*(\lambda^\intercal h(Y,b))|] < \infty\right\} = \left\{\lambda \in \mathbb{R}^2 \; ; \; \int|\exp(\lambda_1(y-b) + \lambda_2) - 1|dy|] < \infty\right\}$. The implied solution to the primal, $q^b(y) = (f^*)'(\lambda^\intercal h(y,b)) = \exp(\lambda_1(y-b) + \lambda_2)$ satisfies $0 < q^b(y) < \infty$ on the support of P_1 and solves the moment conditions. Thus assumption 2 is satisfied for any $B \subset \left(\frac{p_D}{2}, 1 - \frac{p_D}{2}\right)$.

A.2 When ν is convex

Lemma 4.3 (Convex value function). *Let B be convex, assumption 1 hold, and $g(y,x,b) = \tilde{g}(y,x) - b$. Then $\hat{\nu}_n$ and ν are convex on B. If assumption 2 holds as well, then ν is strictly convex on B.*

Proof. Let $b^0, b^1 \in B$, $\alpha \in (0,1)$, and $b^\alpha = \alpha b^1 + (1-\alpha)b^0$. The proof will show

$$\alpha\nu(b^1) + (1-\alpha)\nu(b^0) \geq \nu(b^\alpha)$$

directly.

Let $\{Q^{0,n}\}_{n=1}^\infty \subset \mathbf{P}^{b^0}$ and $\{Q^{1,n}\}_{n=1}^\infty \subset \mathbf{P}^{b^1}$ be such that

$$d_f(Q^{0,n} \| P_1) \to \inf_{Q \in \mathbf{P}^{b^0}} d_f(Q\|P_1) = \nu(b_0), \qquad d_f(Q^{1,n} \| P_1) \to \inf_{Q \in \mathbf{P}^{b^1}} d_f(Q\|P_1) = \nu(b_1)$$

Notice that $\{Q^{0,n}\}_{n=1}^{\infty} \subset \mathbf{P}^{b^0}$ and $\{Q^{1,n}\}_{n=1}^{\infty} \subset \mathbf{P}^{b^1}$ implies

$$E_{Q^{1,n}}[\tilde{g}(Y,X)] - b^1 = \frac{-p_D}{1-p_D} E_{P_1}[\tilde{g}(Y,X) - b^1],$$
$$E_{Q^{0,n}}[\tilde{g}(Y,X)] - b^0 = \frac{-p_D}{1-p_D} E_{P_1}[\tilde{g}(Y,X) - b^0]$$

Which implies that

$$E_{\alpha Q^{1,n}+(1-\alpha)Q^{0,n}}[\tilde{g}(Y,X)] - (\alpha b^1 + (1-\alpha)b^0) = \frac{-p_D}{1-p_D} E_{P_1}\left[\tilde{g}(Y,X) - (\alpha b^1 + (1-\alpha)b^0)\right]$$

Similarly, $E_{Q^{0,n}}[\mathbb{1}\{X = x_k\}] = E_{Q^{1,n}}[\mathbb{1}\{X = x_k\}] = E_{P_{0X}}[\mathbb{1}\{X = x_k\}]$ for all $k = 1, \ldots, K$. It follows that $Q^{\alpha,n} := \alpha Q^{1,n} + (1-\alpha)Q^{0,n}$ is feasible for $b^\alpha = \alpha b^1 + (1-\alpha)b^0$. This implies

$$d_f(Q^{\alpha,n} \| P_1) \geq \inf_{Q \in \mathbf{P}^{b^\alpha}} d_f(Q \| P_1) = \nu(b^\alpha)$$

$\{Q^{0,n}\}_{n=1}^{\infty}$ and $\{Q^{1,n}\}_{n=1}^{\infty}$ have densities with respect to Q, denoted $\{q^{0,n}\}_{n=1}^{\infty}$ and $\{q^{1,n}\}_{n=1}^{\infty}$ respectively. Convexity of f implies that for any (y,x),

$$\alpha f(q^{1,n}(y,x)) + (1-\alpha) f(q^{0,n}(y,x)) \geq f(\alpha q^{1,n}(y,x) + (1-\alpha) q^{0,n}(y,x))$$

integrating with respect to P_1 shows that

$$\alpha d_f(Q^{1,n} \| P_1) + (1-\alpha) d_f(Q^{0,n} \| P_1) \geq d_f(Q^{\alpha,n} \| P_1) \geq \nu(b^\alpha)$$

Letting $n \to \infty$ gives the result that $\nu(b)$ is convex in b.

Next, notice that no properties of P_1, P_{0X} were specified in the argument above, so the same argument works to show $\hat{\nu}_n(b)$ is convex in b by replacing P_1, P_{0X} with their empirical counterparts.

Finally, suppose further that assumptions 1 and 2 hold, and that $b^0 \neq b^1$. The proof of theorem (3.1) shows that the primal problem at b^0 and b^1 is attained by Q^0 and Q^1 with densities q^0, q^1. Since $E_{Q^0}[\tilde{g}(X,Y)] \neq E_{Q^1}[\tilde{g}(X,Y)]$, the densities q^0, q^1 must differ on a P_1 non-neglible set. For (y,x) in that set, strict convexity of f assumed in (1) (iv) implies

$$\alpha f(q^1(y,x)) + (1-\alpha) f(q^0(y,x)) > f(\alpha q^1(y,x) + (1-\alpha) q^0(y,x))$$

integrating with respect to P_1 gives $\alpha d_f(Q^1 \| P_1) + (1-\alpha) d_f(Q^0 \| P_1) > d_f(\alpha Q^1 + (1-\alpha) Q^0 \| P_1)$, equivalently, $\alpha \nu(b^1) + (1-\alpha)\nu(b^0) > \nu(\alpha b^1 + (1-\alpha) b^0)$.

□

Lemma 4.4 (Convex value function, linear models)**.** *Let B be convex and assumption 1 hold. Suppose the sample is $\{D_i, D_i Y_i, X_{i1}, X_{i2}\}_{i=1}^n$, where $Y_i \in \mathbb{R}$, $X_{i1} \in \mathbb{R}^{d_{x1}}$, and $X_{i2} \in \mathbb{R}^{d_{x2}}$ with $d_{x2} \geq d_{x1}$. Consider an instrumental variables model:*

$$Y_i = X_{i1}^{\intercal} \beta + \varepsilon, \qquad E\left[X_{i2} \varepsilon\right] = 0$$

Then $\hat{\nu}_n$ and ν are convex on B.

Proof. The proof is almost identical to that of Lemma 4.3.

Let $b^0, b^1 \in B$, $\alpha \in (0,1)$, and $b^\alpha = \alpha b^1 + (1-\alpha)b^0$. Let $\{Q^{0,n}\}_{n=1}^\infty \subset \mathbf{P}^{b^0}$ and $\{Q^{1,n}\}_{n=1}^\infty \subset \mathbf{P}^{b^1}$ be such that

$$d_f(Q^{0,n} \| P_1) \to \inf_{Q \in \mathbf{P}^{b^0}} d_f(Q \| P_1) = \nu(b_0), \qquad d_f(Q^{1,n} \| P_1) \to \inf_{Q \in \mathbf{P}^{b^1}} d_f(Q \| P_1) = \nu(b_1)$$

Notice the moment conditions are $0 = E\left[X_2(Y - X_1^\intercal \beta)\right] = E\left[X_2 Y\right] - E\left[X_2 X_1^\intercal\right]\beta$. So $\{Q^{0,n}\}_{n=1}^\infty \subset \mathbf{P}^{b^0}$ and $\{Q^{1,n}\}_{n=1}^\infty \subset \mathbf{P}^{b^1}$ implies that

$$E_{Q^{1,n}}\left[X_2 Y\right] - E_{P_{0X}}\left[X_2 X_1^\intercal\right] b^1 = \frac{-p_D}{1-p_D}\left(E_{P_1}\left[X_2 Y\right] - E_{P_1}\left[X_1 X^\intercal\right] b^1\right),$$
$$E_{Q^{0,n}}\left[X_2 Y\right] - E_{P_{0X}}\left[X_2 X_1^\intercal\right] b^0 = \frac{-p_D}{1-p_D}\left(E_{P_1}\left[X_2 Y\right] - E_{P_1}\left[X_2 X_1^\intercal\right] b^0\right)$$

implying that

$$E_{\alpha Q^{1,n} + (1-\alpha)Q^{0,n}}\left[X_2 Y\right] - E_{P_{0X}}\left[X_2 X_1^\intercal\right] b^\alpha = \frac{-p_D}{1-p_D}\left(E_{P_1}\left[X_2 Y\right] - E_{P_1}\left[X_2 X_1^\intercal\right] b^\alpha\right)$$

Similarly, $E_{Q^{0,n}}[\mathbb{1}\{X = x_k\}] = E_{Q^{1,n}}[\mathbb{1}\{X = x_k\}] = E_{P_{0X}}[\mathbb{1}\{X = x_k\}]$ for all $k = 1, \ldots, K$. It follows that $Q^{\alpha,n} := \alpha Q^{1,n} + (1-\alpha)Q^{0,n}$ is feasible for $b^\alpha = \alpha b^1 + (1-\alpha)b^0$. This implies

$$d_f(Q^{\alpha,n} \| P_1) \geq \inf_{Q \in \mathbf{P}^{b^\alpha}} d_f(Q \| P_1) = \nu(b^\alpha)$$

$\{Q^{0,n}\}_{n=1}^\infty$ and $\{Q^{1,n}\}_{n=1}^\infty$ have densities with respect to Q, denoted $\{q^{0,n}\}_{n=1}^\infty$ and $\{q^{1,n}\}_{n=1}^\infty$ respectively. Convexity of f implies that for any (y,x),

$$\alpha f(q^{1,n}(y,x)) + (1-\alpha) f(q^{0,n}(y,x)) \geq f(\alpha q^{1,n}(y,x) + (1-\alpha) q^{0,n}(y,x))$$

integrating with respect to P_1 shows that

$$\alpha d_f(Q^{1,n} \| P_1) + (1-\alpha) d_f(Q^{0,n} \| P_1) \geq d_f(Q^{\alpha,n} \| P_1) \geq \nu(b^\alpha)$$

Letting $n \to \infty$ gives the result that $\nu(b)$ is convex in b.

Finally, notice that no properties of P_1, P_{0X} were specified in the argument above, so the same argument works to show $\hat{\nu}_n(b)$ is convex in b by replacing P_1, P_{0X} with their empirical counterparts. □

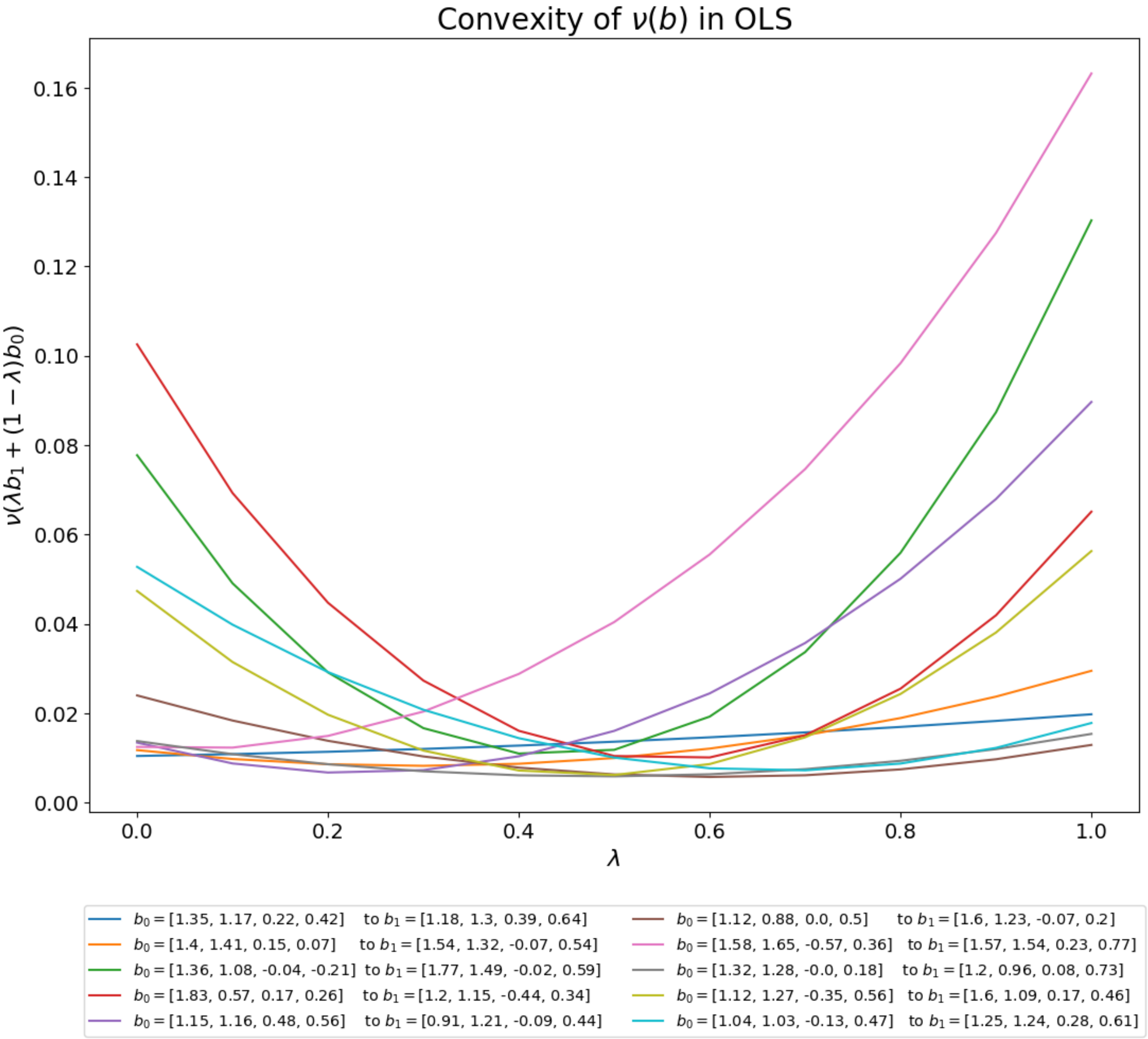
Simulations suggest that OLS more generally produces convex $\nu(b)$. Consider the data generating process described in section 5.2. Here the data is of the form $\{D_i, D_i Y_{i1}, D_i Y_{i2}, X_{i1}, X_{i2}\}_{i=1}^n$, and the model is given by

$$Y_{i1} = \beta_0 + \beta_1 X_1 + \beta_2 Y_2 + \beta_3 X_2 + \varepsilon, \qquad E\left[\begin{pmatrix} 1 \\ X_1 \\ Y_2 \\ X_2 \end{pmatrix} \varepsilon\right] = 0$$

The following figure investigates convexity of the $\nu(b)$ (where $d_f(Q\|P) = H^2(Q,P)$) numerically, by looking for convexity along random line segments. Specifically, let b_1 and b_0 be points in the sample space and $b^\lambda = \lambda b_1 + (1-\lambda) b_0$. Then compute $\hat{\nu}_n(b^\lambda)$ for $\lambda \in \{0, 0.1, 0.2, \ldots, 1\}$.[15] The

[15] Note that while $\hat{\nu}_n(b)$ is the estimator rather than the population value of the DGP described in 5.2, it is the

following figure shows the results of this exercise for 10 randomly selected (b_0, b_1) pairs.



population value of the discretely supported DGP placing uniform mass on the sample's observations.

B Measuring selection and breakdown analysis

B.1 Measuring selection

Lemma 2.1 is found in subsection 2.1.

Lemma 2.1. *Let* $(Z, D) \in \mathbb{R}^{d_z} \times \{0,1\}$ *be random variables with* $p_D = P(D = 1) \in (0,1)$. *Let* $Z \mid D = 1 \sim P_1$ *and* $Z \mid D = 0 \sim P_0$. *Then*

$$H^2(P_0, P_1) = 1 - \frac{E\left[\sqrt{Var(D \mid Z)}\right]}{\sqrt{Var(D)}} \tag{1}$$

where the expectation is taken with respect to $p_D P_1 + (1 - p_D)P_0$, *the marginal distribution of* Z.

Proof. The marginal, unconditional distribution of Z is $P = p_D P_1 + (1 - p_D)P_0$. This distribution dominates P_1 and P_0, which have densities

$$f_1(z) = \frac{P(D = 1 \mid Z = z)}{p_D}, \qquad f_0(z) = \frac{(1 - P(D = 1 \mid Z = z))}{1 - p_D},$$

with respect to P. This implies

$$\begin{aligned} H^2(P_0, P_1) &= \frac{1}{2}\int\left(\sqrt{f_0(z)} - \sqrt{f_1(z)}\right)^2 dP(z) = \frac{1}{2}\left[\int f_0(z) + f_1(z) - 2\sqrt{f_1(z)f_0(z)}dP(z)\right] \\ &= 1 - \frac{\int\sqrt{P(D = 1 \mid Z = z)(1 - P(D = 1 \mid Z = z))}dP(z)}{\sqrt{p_D(1 - p_D)}} \\ &= 1 - \frac{E_P\left[\sqrt{\text{Var}(D \mid Z)}\right]}{\sqrt{\text{Var}(D)}}. \end{aligned}$$

□

B.2 Nominally identified sets

The exercise proposed in section 2.3 can also be understood with a framework of nominally identified sets. This approach to exposition is used in Kline & Santos (2013), Masten & Poirier (2020), and Diegert et al. (2022), and described for the current setting in this appendix.

Under the assumption $d(P_0 \| P_1) \leq \delta$ and $P_0 \ll P_1$, the identified set for β_P is a function of δ:

$$\mathbf{B}_{ID}(\delta) = \left\{b \in \boldsymbol{B}\ ;\ \exists Q,\ \ p_D\mathbb{E}_{P_1}[g(Z, b)] + (1 - p_D)\mathbb{E}_Q[g(Z, b)] = 0,\ \text{and}\ \ d(Q \parallel P_1) \leq \delta\right\} \tag{20}$$

Notice $\mathbf{B}_{ID}(\delta)$ is always growing with δ, in the sense that $\delta < \delta' \implies \mathbf{B}_{ID}(\delta) \subseteq \mathbf{B}_{ID}(\delta')$.

The researcher is primarily interested in testing $H_0 : \beta \in \mathbf{B}_0$ against $H_1 : \beta \in \mathbf{B}_1 = \mathbf{B} \setminus \mathbf{B}_0$. Naturally, if $\mathbf{B}_{ID}(\delta)$ has trivial intersection with $\mathbf{B}_0$ she is confident in rejecting H_0. This leads to the question "what is the largest value of δ such that $\mathbf{B}_{ID}(\delta)$ has empty intersection with $\mathbf{B}_0$?" Formally, define the **breakdown point** as

$$\bar{\delta}^{BD} = \sup\{\delta \in \mathbb{R}_+\ ;\ \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 = \varnothing\} \tag{21}$$

if $\mathbf{B}_{ID}(0) \cap \mathbf{B}_0 = \varnothing$, otherwise define $\bar{\delta}^{BD} := 0$.

B.2.1 Characterization through a value function

Let

$$\mathbf{P}^b = \{Q \; ; \; Q \ll P_1, \; Q_X = P_{0X}, \; p_D E_{P_1}[g(Z,b)] + (1-p_D)E_Q[g(Z,b)] = 0\},$$

be the set of distributions that "rationalizes" $\beta = b$. Notice that if there exists $Q \in \mathbf{P}^b$ such that $d(Q \parallel P_1) \leq \delta$, then $b \in \boldsymbol{B}_{ID}(\delta)$. This suggests the identified sets can be characterized through the value function

$$\nu(b) = \inf_{Q \in \mathbf{P}_b(Q)} d(Q \parallel P_1), \tag{22}$$

where the infimum over the empty set is defined to be $+\infty$. Observe that $\nu(b) < \delta$ implies $b \in \mathbf{B}_{ID}(\delta)$, and if the infimum is attained at some minimum, then $\nu(b) \leq \delta$ if and only if $b \in \mathbf{B}_{ID}(\delta)$.

Lemma B.1 shows that the definition of the breakdown point given in (21) is equivalent to that given by (4).

Lemma B.1 (Characterization of breakdown point)**.**

$$\inf_{b \in \boldsymbol{B}_0} \nu(b) = \bar{\delta}^{BD}$$

Proof. Define the "robust region" as the set of $\delta \in \mathbb{R}_+$ where the identified set has trivial intersection with the null hypothesis:

$$RR = \{\delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 = \varnothing\}$$

and let $RR^c = \mathbb{R}_+ \setminus RR = \{\delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\}$ be its compliment in $\mathbb{R}_+$. Notice that

$$\bar{\delta}^{BD} = \begin{cases} \sup RR & \text{if } RR \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

The proof consists of two steps:

1. Showing that
$$\bar{\delta}^{BD} = \inf RR^c \tag{23}$$
where the infimum over the empty set is defined to be ∞.

2. Arguing that
$$\inf_{b \in \boldsymbol{B}_0} \nu(b) \leq \inf RR^c, \qquad \text{and} \qquad \inf_{b \in \boldsymbol{B}_0} \nu(b) \geq \inf RR^c,$$

Step 1. is a consequence of $\mathbf{B}_{ID}(\delta)$ being a growing set (in the sense that $\delta \leq \delta' \implies \mathbf{B}_{ID}(\delta) \subseteq \mathbf{B}_{ID}(\delta')$). Define $\bar{\delta}^* = \inf RR^c = \inf\{\delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\}$. There are three possibilities:

(i) $\delta^{BD} = 0$. Then RR^c contains $(0, \infty)$, hence $0 \leq \bar{\delta}^* = \inf RR^c \leq \inf(0,\infty) = 0$.

(ii) $\delta^{BD} \in (0, \infty)$. Notice that $\delta \leq \delta' \implies \mathbf{B}_{ID}(\delta) \subseteq \mathbf{B}_{ID}(\delta')$ implies that $\delta \leq \delta' \implies (\mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0) \subseteq (\mathbf{B}_{ID}(\delta') \cap \mathbf{B}_0)$, from which it follows that

$$\begin{aligned} \delta \leq \delta' \text{ and } \delta' \in RR \qquad &\implies \qquad \delta \in RR \\ \delta \leq \delta' \text{ and } \delta \in RR^c \qquad &\implies \qquad \delta' \in RR^c \end{aligned}$$

since $\bar{\delta}^{BD} \in (0, \infty)$, we have RR contains $[0, \bar{\delta}^{BD})$. Similarly, RR^c contains $(\bar{\delta}^*, \infty)$, and since $RR \cap RR^c = \varnothing$, we have $\bar{\delta}^{BD} \leq \bar{\delta}^*$. For $n \in \mathbb{N}$, let $\delta_n := \bar{\delta}^* - \frac{1}{n} \geq 0$, and notice that $\mathbf{B}_{ID}(\delta_n) \cap B = \varnothing$, equivalently, $\delta_n \in RR$. Therefore

$$\bar{\delta}^* - \frac{1}{n} \leq \bar{\delta}^{BD} \leq \bar{\delta}^*$$

let $n \to \infty$ to see that $\delta^{BD} = \delta^*$.

(iii) $\delta^{BD} = \infty$. Then the argument above implies RR contains $[0, \infty)$, so $RR^c = \varnothing$ and $\delta^* = \infty$.

Therefore (23) holds.

For step 2., first notice that

$$\inf_{b \in \boldsymbol{B}_0} \nu(b) = \inf_{b \in \boldsymbol{B}_0} \inf_{Q \in \mathbf{P}^b} d(Q \parallel P_1) = \inf \bigcup_{b \in \boldsymbol{B}_0} \left\{ d(Q \parallel P_1) \; ; \; Q \in \mathbf{P}^b \right\} \tag{24}$$

If δ is such that $\mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing$, then there exists $b \in \mathbf{B}_0$ and $Q \in \mathbf{P}^b$ such that $d(Q \parallel P_1) \leq \delta$. This implies

$$\inf RR^c = \inf \{\delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\} \geq \inf \bigcup_{b \in \mathbf{B}_0} \left\{ d(Q \parallel P_1) \; ; \; K \in \mathbf{P}^b \right\}$$

Conversely, for each real number a satisfying $a = d(Q \parallel P_1)$ for some $Q \in \mathbf{P}^b$, $b \in \mathbf{B}_0$, we have that $a \in \{\delta \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\}$. This implies

$$\inf \bigcup_{b \in \mathbf{B}_0} \{d(Q \parallel P_1) \; ; \; Q \in \mathbf{P}_b\} \geq \inf \{\delta \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\} = \inf RR^c$$

Putting (23), (24), and these two inequalities together we obtain

$$\inf_{b \in \mathbf{B}_0} \nu(b) = \inf \bigcup_{b \in \mathbf{B}_0} \left\{ d(Q \parallel P_1) \; ; \; Q \in \mathbf{P}^b \right\} = \inf \{\delta \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\} = \delta^{BD}$$

as was claimed. □ □

C Additional duality discussion

This appendix contains no original results, but collects useful definitions and facts from convex analysis.

C.1 Definitions

For reference, see Broniatowski & Keziou (2012), or Rockafellar (1970).

Let $f : \mathbb{R} \to (-\infty, \infty]$. The *effective domain* of f is $\mathrm{dom}(f) = \{x \in \mathbb{R} \; ; \; f(x) < \infty\}$. f is called *proper* if $\mathrm{dom}(f)$ is nonempty. f is called *convex* if $\mathrm{dom}(f)$ is a convex set. For a convex $f : E \subsetneq \mathbb{R} \to \mathbb{R}$, we can extend f to $\mathbb{R}$ by setting $f(x) = \infty$ for all $x \notin E$. This extended function is still convex.

Now consider a convex $f : \mathbb{R} \to (-\infty, \infty]$. Notice that convexity implies $\mathrm{dom}(f)$ is a subset of $\mathbb{R}$ with interior of the form (ℓ, u). ℓ or u may be infinite, and $\lim_{x \rightarrowtail \ell^+} f(x)$ or $\lim_{x \to u^-} f(u)$ may be finite. f is called *closed* if

1. $\lim_{x\to\ell^+} f(x) = -\infty$ if $\ell > -\infty$, and
2. $\lim_{x\to u^-} f(x) = \infty$ if $u < \infty$.

f is called *essentially smooth* if 1. f is differentiable on (ℓ, u), 2. $\lim_{x\to\ell^+} f'(x) = -\infty$ if $\ell > -\infty$, and 3. $\lim_{x\to u^-} f'(x) = \infty$ if $u < \infty$. The *convex conjugate* or *Legendre-Fenchel transform* of a convex function f is defined as $f^*(y) = \sup_{x\in\mathbb{R}}\{xy - f(x)\}$.

C.2 Results

Now let f be closed, proper, and convex. The following are results not proven here; see footnotes for references.

- f^* is a closed, proper, convex function.[16]
- $(f^*)^* = f$; that is, the convex conjugate of f^* is f.[17]
- f is strictly convex if and only if f^* is essentially smooth.[18]
- f is essentially smooth if and only if f^* is strictly convex.[19]
- If f is strictly convex and essentially smooth, then f' is one-to-one and $(f')^{-1}(y) = (f^*)'(y)$ for all $y \in \mathrm{dom}(f^*)$.[20]
- If f is strictly convex, essentially smooth, and twice differentiable, then f^* is twice differentiable and $(f^*)''(y) = \frac{1}{f''((f')^{-1}(y))}$.[21]
- If f is strictly convex and essentially smooth with $\mathrm{dom}(f) \subseteq [0,\infty)$, then $(f')^{-1}(x) \geq 0$.[22]
- If f is convex, $f(x) = 0$ at $x = 1$, and f is strictly convex on a neighborhood of 1 then $\int f(k(z))dP(z) = 0$ if and only if $k(z) = 1$, P-a.s..[23]

D Duality Results

Lemma D.1 (Unique primal solution)**.** *Suppose f is strictly convex on its domain and $p_D \in (0,1)$. Then any solution attaining the infimum in* (5) *is unique, P_1-almost surely.*

Proof. Let $Q^0, Q^1 \in \mathbf{P}^b$ attain the infimum in (4), and let q^0 and q^1 denote their densities with respect to P_1. We have that $\frac{-p_D}{(1-p_D)}E_{P_1}[g(Z,b)] = E_{Q^0}[g(Z,b)] = E_{Q^1}[g(Z,b)]$ and $Q^0_X = Q^1_X = P_{0X}$. For any $\alpha \in (0,1)$, the measure $Q^\alpha = \alpha Q^1 + (1-\alpha)Q^0 \in \mathbf{P}^b$ is feasible in (5), and characterized by the P_1-density $\alpha q^1 + (1-\alpha)q^0$.

[16] Rockafellar (1970), p. 104.

[17] Rockafellar (1970) p. 104, theorem 12.2

[18] Borwein & Lewis (1993) p. 251, or Rockafellar (1970) theorem 26.3 on p. 253.

[19] This follows from the two preceding facts.

[20] Broniatowski & Keziou (2012) p. 2559. See also Rockafellar (1970) corollary 23.5.1 on p. 219, corollary 26.3.1 on p. 254, and theorem 26.5 on p. 258. For a sketch of the proof, let $\phi(x,y) = xy - f(x)$ and notice $f^*(y) = \sup_x \phi(x,y)$. The FOC implies the optimal x is given by $\tilde{x}(y) = (f')^{-1}(y)$, hence $f^*(y) = \phi(\tilde{x}(y), y)$. Apply the envelope theorem to find $\frac{d}{dy}f^*(y) = \underbrace{\frac{\partial\phi}{\partial x}(\tilde{x}(y),y)}_{=0}\tilde{x}'(y) + \frac{\partial\phi}{\partial y}(\tilde{x}(y),y) = \frac{\partial\phi}{\partial x}(\tilde{x}(y),x) = \tilde{x}(y) = (f')^{-1}(y)$.

[21] Broniatowski & Keziou (2012), p. 2559. See also the preceding fact.

[22] Broniatowski & Keziou (2012), p. 2557.

[23] Broniatowski & Keziou (2012), p. 2556.

Suppose for contradiction that Q^0 and Q^1 differ on a set of positive P_1-measure. Strict convexity implies that for any z in that set,

$$f(\alpha q^1(z) + (1-\alpha)q^0(z)) < \alpha f(q^1(z)) + (1-\alpha)q^0(z)$$

Integrating with respect to P_1 reveals $d_f(\alpha Q^1 + (1-\alpha)Q^0 \| P_1) < \alpha d_f(Q^1\|P_1) + (1-\alpha)d_f(Q^0\|P_1)$, contradicting optimality of Q^0, Q^1. □

Lemma D.2 (Weak duality)**.** *Let $\nu(b)$ and $V(b)$ be as defined in* (5) *and* (7)*, respectively. If assumption 1 holds, then b, $V(b) \leq \nu(b)$ for any $b \in \boldsymbol{B}$.*

Proof. First note that if $\nu(b) = \infty$ the inequality holds trivially.

Suppose $\nu(b) < \infty$. Then $\mathbf{P}^b \neq \varnothing$, hence there exists at least one density $q(z) = \frac{dQ}{dP_1}(z)$ satisfying $\int h(z,b)q(z)dP_1(z) = c(b)$. Notice that $f^*(r) = \sup_{t\in\mathbb{R}}\{rt - f(t)\}$ implies $f(t) + f^*(r) \geq f(t) + rt - f(t) = rt$. Apply this to find that for any $Q \in \mathbf{P}^b$ with P_1-density q,

$$\begin{aligned} & f(q(z)) + f^*(\lambda^\intercal h(z,b)) \geq \lambda^\intercal h(z,b)q(z) \\ & \qquad \implies f(q(z)) \geq \lambda^\intercal h(z,b)q(z) - f^*(\lambda^\intercal h(z,b)) \end{aligned}$$

integrating over z with respect to P_1 gives

$$\begin{aligned} & \int f(q(z))dP_1(z) \geq \lambda^\intercal \underbrace{\int h(z,b)q(z)dP_1(z)}_{=c(b)} - \int f^*(\lambda^\intercal h(z,b))dP_1(z) \\ & \implies d_f(Q\|P_1) \geq \lambda^\intercal c(b) - E\left[f^*(\lambda^\intercal h(z,b)) \mid D=1\right] \end{aligned}$$

the left hand side of the last inequality doesn't depend on $\lambda \in \mathbb{R}^{d_g+K}$, while the right hand side doesn't depend on $Q \in \mathbf{P}^b$. Hence,

$$\nu(b) = \inf_{Q\in\mathbf{P}^b} d_f(Q\|P_1) \geq \sup_{\lambda\in\mathbb{R}^{d_g+K}} \left\{\lambda^\intercal c(b) - E\left[f^*(\lambda^\intercal h(Z,b)) \mid D=1\right]\right\} = V(b)$$

□

Theorem 3.1 is found in section 3.1.

Theorem 3.1 (Strong duality)**.** *Suppose assumptions 1 and 2 hold. Then for each $b \in B$, $\nu(b) = V(b)$, with dual attainment.*

Proof. Let $\mathbf{M}$ be the set of measurable functions mapping $z = (x,y) \mapsto \mathbb{R}$. Consider the relaxed problem

$$\begin{aligned} \tilde{\nu}(b) &= \inf_{q\in\tilde{\mathbf{P}}^b} \int f(q(z))dP_1(z) \\ \tilde{\mathbf{P}}^b &= \Bigg\{ q \in \mathbf{M}\,;\, \int \mathbb{1}\{x = x_k\}q(x,y)dP_1(x,y) = P(X = x_k \mid D=0) \text{ for } k = 1,\ldots,K, \\ & \qquad\qquad\qquad\qquad \text{and} \quad \int g(z,b)q(z)dP_1(z) = c(b) \Bigg\} \end{aligned}$$

for any $q \in \tilde{\mathbf{P}}^b$, $K(\psi) = \int \psi(z)q(z)dP_1(z)$ is a (possibly signed) measure with total measure one. Notice this problem has the same objective as the primal problem (5), but allows for finite signed measures.

Now apply Theorem II.2 of Csiszár et al. (1999). The dual of the relaxed problem is (7). Assumption 2 (i) is the "constraint qualification" of Csiszár et al. (1999) Theorem II.2, implying strong duality holds for the relaxed problem, $\tilde{\nu}(b) = V(b)$, and the dual problem's value is attained at a maximum. Let $\lambda(b)$ solve the dual problem. Assumption 2 (ii) allows application of the second part of Theorem II.2, implying the solution to the relaxed problem is given by

$$\begin{aligned} q^b(z) &= (f')^{-1}(\lambda(b)^\intercal h(z,b)) \\ &= (f^*)'(\lambda(b)^\intercal h(z,b)) \end{aligned}$$

By assumption 1 (iv) and Lemma D.1, this solution is unique P_1-almost surely.

Now we show that q^b in fact solves the primal problem, (5). Notice q^b is nonnegative, because f' is only defined on the non-negative reals. Furthermore,

$$\begin{aligned} \int q(x,y)dP_1(x,y) &= \int \sum_{k=1}^{K} \mathbb{1}\{x = x_k\}q(x,y)dP_1(x,y) \\ &= \sum_{k=1}^{K} \int \mathbb{1}\{x = x_k\}q(x,y)dP_1(x,y) \\ &= \sum_{k=1}^{K} P(X = x_k \mid D = 0) \\ &= 1 \end{aligned}$$

So the measure Q^b given by $Q^b(\psi) = \int \psi(z)q^b(z)dP_1(z)$ is a probability distribution dominated by P_1. Therefore $Q^b \in \mathbf{P}^b$ is feasible in the primal problem (5). Being feasible in the primal and solving the relaxed problem, Q^b must also solve the primal problem. □

Remark D.1. See also Broniatowski & Keziou (2006), theorem 1.

E Technical Lemmas

These results are self contained, with notation not related to the present paper.

For any set $\mathcal{X}$, $\ell^\infty(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \ ; \ \sup_{x\in\mathcal{X}}|f(x)| < \infty\}$ denotes the set of bounded functions on $\mathcal{X}$. $\ell^\infty(\mathcal{X})$ is equipped with the sup-norm: for $f \in \ell^\infty(\mathcal{X})$, $\|f\|_\infty = \|f\|_{\mathcal{X}} = \sup_{x\in\mathcal{X}}|f(x)|$.

E.1 Technical lemmas used to show consistency

Lemma E.1 (Glivenko-Cantelli with Nuisance Parameter)**.** *Let $x \in \mathbb{R}^{d_x}$, $\theta \in \mathbb{R}^{d_\theta}$, $\gamma \in \mathbb{R}^{d_\gamma}$, and $f(x,\theta,\gamma) \in \mathbb{R}^{d_f}$ all be finite dimensional. If*

(i) $\{X_i\}_{i=1}^n$ are i.i.d.

(ii) $\hat{\gamma}_n \xrightarrow{p} \gamma_0$,

(iii) $a(x,\gamma) := \sup_{\theta\in\Theta} \|f(x,\theta,\gamma) - f(x,\theta,\gamma_0)\|$ is continuous in γ at γ_0

(iv) there exists an open set $\mathcal{N} \subset \mathbb{R}^{d_\gamma}$ satisfying $E\left[\sup_{\gamma\in\mathcal{N}} \sup_{\theta\in\Theta} \|f(X,\theta,\gamma)\|\right] < \infty$ and

(v) $\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n f(X_i,\theta,\gamma_0) - E\left[f(X,\theta,\gamma_0)\right]\right\| \xrightarrow{p} 0$,

then $\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n f(X_i,\theta,\hat{\gamma}_n) - E[f(X,\theta,\gamma_0)]\right\| \xrightarrow{p} 0$.

Proof. Condition (v) and the triangle inequality implies

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n f(X_i,\theta,\hat{\gamma}_n) - E[f(X,\theta,\gamma_0)]\right\| \\
&\quad \leq \sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n f(X_i,\theta,\hat{\gamma}_n) - \frac{1}{n}\sum_{i=1}^n f(X_i,\theta,\gamma_0)\right\| + \underbrace{\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n f(X_i,\theta,\gamma_0) - E[f(X,\theta,\gamma_0)]\right\|}_{=o_p(1) \text{ by condition } (v)} \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n \sup_{\theta\in\Theta} \underbrace{\|f(X_i,\theta,\hat{\gamma}_n) - f(X_i,\theta,\gamma_0)\|}_{=a(X_i,\hat{\gamma}_n)} + o_p(1)
\end{aligned}$$

so it suffices to show $\frac{1}{n}\sum_{i=1}^n a(X_i,\hat{\gamma}_n) \xrightarrow{p} 0$.

$\hat{\gamma}_n \xrightarrow{p} \gamma_0$ implies there exists $\delta_n \to 0$ such that $\|\hat{\gamma}_n - \gamma_0\| \leq \delta_n$ with probability approaching one. Define $\Delta_n(x) = \sup_{\gamma\,;\,\|\gamma-\gamma_0\|\leq\delta_n} a(x,\gamma)$. Since $a(x,\gamma_0) = 0$ and $a(x,\cdot)$ is continuous at γ_0 by condition (iii), $\Delta_n(x) \to 0$ pointwise. For large enough n, $\{\gamma\,;\,\|\gamma-\gamma_0\|\leq\delta_n\} \subset \mathcal{N}$ for the $\mathcal{N}$ from condition (iv) and hence $\Delta_n(x) \leq 2\sup_{\lambda\in\mathcal{N}}\sup_{\theta\in\Theta}\|f(x,\theta,\gamma)\|$ eventually. The dominated convergence theorem then implies $E[\Delta_n(X)] \to 0$. Now apply Markov's inequality and condition (i) to find

$$P\left(\frac{1}{n}\sum_{i=1}^n \Delta_n(X_i) > \varepsilon\right) \leq \frac{E[\Delta_n(X)]}{\varepsilon} \to 0$$

therefore $\frac{1}{n}\sum_{i=1}^n \Delta_n(X_i) \xrightarrow{p} 0$. Finally, with probability approaching one $\|\hat{\gamma}_n - \gamma_0\| \leq \delta_n$ and on that event,

$$\frac{1}{n}\sum_{i=1}^n a(X_i,\hat{\gamma}_n) \leq \frac{1}{n}\sum_{i=1}^n \Delta_n(X_i) \xrightarrow{p} 0$$

which completes the proof. □

Remark E.1. The proof of lemma E.1 is quite similar to that of lemma 4.3 in Newey & McFadden (1994).

Lemma E.2 (Restricted infimum is continuous)**.** *For any bounded $f, g : X \to \mathbb{R}$ and any $D \subseteq X$,*

$$\left|\inf_{x\in D} f(x) - \inf_{x\in D} g(x)\right| \leq \sup_{x\in D}|f(x) - g(x)|$$

as a result, $\iota_D : \ell^\infty(X) \to \mathbb{R}$ given by $\iota_D(h) = \inf_{x\in D} h(x)$ is continuous.

Proof. Let $h, k \in \ell^\infty(X)$ be arbitrary, and notice that

$$\begin{aligned}
\sup_{x\in D} h(x) - \sup_{x\in D} k(x) &\leq \sup_{x\in D}\{h(x) - k(x)\} \leq \sup_{x\in D}|h(x) - k(x)|, \text{ and} \\
-\left[\sup_{x\in D} h(x) - \sup_{x\in D} k(x)\right] = \sup_{x\in D} k(x) - \sup_{x\in D} h(x) &\leq \sup_{x\in D}\{k(x) - h(x)\} \leq \sup_{x\in D}|k(x) - h(x)| = \sup_{x\in D}|h(x) - k(x)|
\end{aligned}$$

hence $-\sup_{x\in D}|h(x)-k(x)| \le \sup_{x\in D} h(x) - \sup_{x\in D} k(x) \le \sup_{x\in D}|h(x)-k(x)|$, or equivalently

$$\left|\sup_{x\in D} h(x) - \sup_{x\in D} k(x)\right| \le \sup_{x\in D}|h(x)-k(x)| \tag{25}$$

Use this to see the claimed inequality:

$$\begin{aligned}\left|\inf_{x\in D} f(x) - \inf_{x\in D} g(x)\right| &= \left|-\sup_{x\in D}\{-f(x)\} - \left(-\sup_{x\in D}\{-g(x)\}\right)\right| \\ &= \left|\sup_{x\in D}\{-g(x)\} - \sup_{x\in D}\{-f(x)\}\right| \\ &\le \sup_{x\in D}|-g(x) - \{-f(x)\}| \\ &= \sup_{x\in D}|f(x)-g(x)|\end{aligned}$$

where the inequality is an application of (25) with $h=-g$ and $k=-f$.

To see the continuity claim, let $\varepsilon>0$ and set $\delta=\varepsilon$. The inequality implies

$$|\iota_D(f)-\iota_D(g)| \le \sup_{x\in D}|f(x)-g(x)| \le \sup_{x\in X}|f(x)-g(x)|$$

hence $\|f-g\|_D \coloneqq \sup_{x\in X}|f(x)-g(x)| < \delta$ implies $|\iota_D(f)-\iota_D(g)|<\varepsilon$. □

E.2 Technical lemmas used in inference

The following lemma gives a sufficient condition for a function to define a continuous map between function spaces. For example, if $f:\mathbb{R}\to\mathbb{R}$, one can define a map $\tilde f:\ell^\infty(T)\to\ell^\infty(T)$ pointwise by $\tilde f(g)(t)=f(g(t))$ and ask when $\tilde f$ is continuous. It is sometimes useful to restrict the domain of $\tilde f$ based on the range of the functions being passed to it, or allow f to depend on t directly as well.

Lemma E.3 (Continuity of maps between bounded functions)**.** *Let $T\subseteq\mathbb{R}^{d_T}$, $E^t\subseteq\mathbb{R}^{d_E}$ for each $t\in T$, $E^T=\{(t,e)\in\mathbb{R}^{d_T}\times\mathbb{R}^{d_E}\ ;\ t\in T,\ e\in E^t\}$, and $\ell^\infty_{E^T}(T)^{d_E}\subset\ell^\infty(T)^{d_E}$ defined by*

$$\ell^\infty_{E^T}(T)^{d_E} = \left\{g:T\to\mathbb{R}^{d_E}\ ;\ g(t)\in E^t,\ \sup_{t\in T}\|g(t)\|<\infty\right\}$$

Let $f:E^T\to\mathbb{R}^{d_f}$, and define $\tilde f$ pointwise:

$$\tilde f:\ell^\infty_{E^T}(T)^{d_E}\to\ell^\infty(T)^{d_f}, \qquad \tilde f(g)(t)=f(t,g(t))$$

If $\{f(t,\cdot)\}_{t\in T}$ is uniformly equicontinuous, then $\tilde f$ is continuous.

Proof. Let $\varepsilon>0$, and choose $\delta>0$ such that

$$|e_1-e_2|<\delta \implies |f(t,e_1)-f(t,e_2)|<\varepsilon/2$$

for any $t\in T$. Notice that if $g_1,g_2\in\ell^\infty_E(T)$ with $\|g_1-g_2\|_T=\sup_{t\in T}|g_1(t)-g_2(t)|<\delta$, then for all $t\in T$, $|g_1(t)-g_2(t)|<\delta$, implying $|f(t,g_1(t)-f(t,g_2(t))|<\varepsilon/2$. It follows that

$$\|\tilde f(g_1)-\tilde f(g_2)\|_T = \sup_{t\in T}|f(t,g_1(t)-f(t,g_2(t))| \le \varepsilon/2<\varepsilon$$

and hence $\|g_1 - g_2\|_T < \delta \implies \|\tilde{f}(g_1) - \tilde{f}(g_2)\|_T < \varepsilon$. □

Remark E.2. Note that if $f : T \times E \to \mathbb{R}^{d_f}$ is uniformly continuous then $\{f(t, \cdot)\}_{t \in T}$ is uniformly equicontinuous.

Remark E.3. Lemma E.3 contains many special cases that will be used below. For example, suppose f does not depend on t, and $E^t = E \subseteq \mathbb{R}$ is the same set for all $t \in T$. Then the result simplifies to: if $f : E \subseteq \mathbb{R} \to \mathbb{R}$ is uniformly continuous, then $\tilde{f} : \ell^\infty(T) \to \ell^\infty(T)$ defined pointwise by $\tilde{f}(g)(t) = f(g(t))$ is continuous.

Lemma E.4 (Uniform consistency of matrix inverses)**.** *Let $\bar{\Phi}_n, \Phi : B \to \mathbb{R}^{K \times K}$, with $\bar{\Phi}_n$ random and Φ deterministic. If*

(i) $\Phi(b)^{-1}$ exists for all $b \in B$,

(ii) $\sup_{b \in B} \|\bar{\Phi}_n(b) - \Phi(b)\|_o \xrightarrow{p} 0$, and

(iii) $\sup_{b \in B} \|\Phi(b)\|_o < \infty$, and $\sup_{b \in B} \|\Phi(b)^{-1}\|_o < \infty$,

then with probability approaching one, the function mapping B to $\bar{\Phi}_n(b)^{-1}$ is well defined and

$$\sup_{b \in B} \left\|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right\|_o \xrightarrow{p} 0$$

Proof. It suffices to show that the singular values of $\bar{\Phi}_n(b)$ converge in probability to the singular values of $\Phi(b)$, uniformly over $b \in B$:

$$\sup_{b \in B} \max_k |\sigma_k(\bar{\Phi}_n(b)) - \sigma_k(\Phi(b))| \xrightarrow{p} 0 \tag{26}$$

To see why, notice that $\infty > \sup_{b \in B} \|\Phi(b)^{-1}\| = \sup_{b \in B} \frac{1}{\sigma_K(\Phi(b))} = \frac{1}{\inf_{b \in B} \sigma_K(\Phi(b))}$ implies $\varepsilon := \inf_{b \in B} \sigma_K(\Phi(b)) > 0$. If (26) holds, then with probability approaching one, $\sup_{b \in B} \max_k |\sigma_k(\bar{\Phi}_n(b)) - \sigma_k(\Phi(b))| < \varepsilon/2$ and on this event the function mapping B to $\bar{\Phi}_n(b)^{-1}$ is well defined. Then notice that

$$\begin{aligned} \|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\|_o &= \left\|\Phi(b)^{-1}\left(\Phi(b) - \bar{\Phi}_n(b)\right)\bar{\Phi}_n(b)^{-1}\right\|_o \\ &\leq \left\|\Phi(b)^{-1}\right\|_o \left\|\Phi(b) - \bar{\Phi}_n(b)\right\|_o \left\|\bar{\Phi}_n(b)^{-1}\right\|_o \end{aligned}$$

implying

$$\sup_{b \in B} \|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\|_o \leq \sup_{b \in B} \left\|\Phi(b)^{-1}\right\|_o \sup_{b \in B} \left\|\bar{\Phi}_n(b) - \Phi(b)\right\|_o \sup_{b \in B} \left\|\bar{\Phi}_n(b)^{-1}\right\|_o \tag{27}$$

$\left\|\Phi(b)^{-1}\right\|_o < \infty$ and $\sup_{b \in B} \|\bar{\Phi}_n(b) - \Phi(b)\|_o \xrightarrow{p} 0$ by assumption, which implies $\sup_{b \in B} \|\bar{\Phi}_n(b)^{-1}\|_o = O_p(1)$ by the continuous mapping theorem. Hence (27) implies $\sup_{b \in B} \|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\|_o \xrightarrow{p} 0$.

The argument that (26) holds is broken into three steps:

1. First, showing that $\bar{\Phi}_n(b)^\intercal \bar{\Phi}_n(b)$ is uniformly consistent for $\Phi(b)^\intercal \Phi(b)$.

Notice that

$$
\begin{aligned}
&\sup_{b\in B} \left\|\bar{\Phi}_n(b)^\intercal\bar{\Phi}_n(b) - \Phi(b)^\intercal\Phi(b)\right\|_o \\
&\qquad \le \sup_{b\in B} \left\|\bar{\Phi}_n(b)^\intercal\bar{\Phi}_n(b) - \bar{\Phi}_n(b)^\intercal\Phi(b)\right\| + \sup_{b\in B} \left\|\bar{\Phi}_n(b)^\intercal\Phi(b) - \Phi(b)^\intercal\Phi(b)\right\| \\
&\qquad \le \sup_{b\in B} \left\|\bar{\Phi}_n(b)^\intercal\right\|_o \sup_{b\in B} \left\|\bar{\Phi}_n(b) - \Phi(b)\right\|_o + \sup_{b\in B} \left\|\bar{\Phi}_n(b)^\intercal - \Phi(b)^\intercal\right\|_o \sup_{b\in B} \|\Phi(b)\|_o
\end{aligned}
$$

Recall that for any matrix $A \in \mathbb{R}^{K\times K}$,

$$\|A^\intercal\|_{\max} = \|A\|_{\max} \le \|A\|_o \le K\|A\|_{\max} = K\|A^\intercal\|_{\max}$$

Which implies $\|\bar{\Phi}_n(b)^\intercal\|_o \le K\|\bar{\Phi}_n(b)\|_o$ and $\|\bar{\Phi}_n(b)^\intercal - \Phi(b)^\intercal\|_o \le K\|\bar{\Phi}_n(b) - \Phi(b)\|_o$, giving us

$$
\begin{aligned}
&\sup_{b\in B} \left\|\bar{\Phi}_n(b)^\intercal\bar{\Phi}_n(b) - \Phi(b)^\intercal\Phi(b)\right\|_o \\
&\qquad \le K \sup_{b\in B} \left\|\bar{\Phi}_n(b) - \Phi(b)\right\|_o \left(\sup_{b\in B} \left\|\bar{\Phi}_n(b)\right\| + \sup_{b\in B} \|\Phi(b)\|_o\right) \qquad (28)
\end{aligned}
$$

$\sup_{b\in B}\|\Phi(b)\|_o < \infty$ and $\sup_{b\in B}\|\bar{\Phi}_n(b)-\Phi(b)\| \xrightarrow{p} 0$ by assumption, which implies $\sup_{b\in B}\|\bar{\Phi}_n(b)\|_o = O_p(1)$ by the continuous mapping theorem. Finally, $\sup_{b\in B} \left\|\bar{\Phi}_n(b)^\intercal\bar{\Phi}_n(b) - \Phi(b)^\intercal\Phi(b)\right\|_o \xrightarrow{p} 0$ by (28).

2. Second, using uniform continuity of the mapping from matrices to eigenvalues to obtain uniform consistency of the square of the singular values, through the continuous mapping theorem.

   Recall Wey's perturbation theorem, found in Bhatia (1997) as corollary III.2.6: for Hermitian matrices A and B,

   $$\max_k |\alpha_k(A) - \alpha_k(B)| \le \|A - B\|_o$$

   For real matrices Hermitian is equivalent to symmetric, so Weyl's result implies

   $$
   \begin{aligned}
   &\sup_{b\in B} \max_k \left|\alpha_k\left(\bar{\Phi}_n(b)^\intercal\bar{\Phi}_n(b)\right) - \alpha_k\left(\Phi(b)^\intercal\Phi(b)\right)\right| \\
   &\qquad \le \sup_{b\in B} \left\|\bar{\Phi}_n^\intercal\bar{\Phi}_n(b) - \Phi(b)^\intercal\Phi(b)\right\|_o \xrightarrow{p} 0
   \end{aligned}
   $$

   In other words, uniform consistency of $\bar{\Phi}_n^\intercal\bar{\Phi}_n(b)$ (argued above) implies uniform consistency of the eigenvalues of $\bar{\Phi}_n^\intercal\bar{\Phi}_n(b)$. These eigenvalues are the squared singular values of $\bar{\Phi}_n(b)$.

3. Third, using uniform continuity of the square root function to complete the proof through the continuous mapping theorem.

   Let $\ell_+^\infty(B)$ denote the set of bounded functions taking nonnegative real values, $h : B \to [0, \infty)$, equipped with the sup norm: $\|h\|_B = \sup_{b\in B}|h(b)|$. Lemma E.3 shows that if $f : [0, \infty) \to \mathbb{R}$ is uniformly continuous, $\tilde{f} : \ell_+^\infty(B) \to \ell^\infty(B)$ given by $\tilde{f}(h)(b) = f(h(b))$ is continuous. It is well known that the square root function $x \mapsto \sqrt{x}$ is uniformly continuous on $[0, \infty)$. (26) follows by the continuous mapping theorem.

Having shown (26), the argument above implies $\bar{\Phi}_n(b)^{-1}$ is well defined with probability approaching one, and $\left\|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right\|_o \xrightarrow{p} 0$.

□

Lemma E.5 (Hadamard directional differentiability of infimum)**.** *Let B be compact, $\boldsymbol{B}_0$ closed, $B \cap \boldsymbol{B}_0 \neq \varnothing$, and*

$$\iota : \ell^\infty(B) \to \mathbb{R}, \qquad\qquad \iota(\nu) = \inf_{b \in B \cap \boldsymbol{B}_0} \nu(b)$$

then ι is Hadamard directionally differentiable tangentially to $\mathcal{C}(B) = \{f : B \to \mathbb{R} \ ; \ f$ is continuous$\}$ at any $\nu \in \mathcal{C}(B)$, with

$$\iota'_\nu(h) = \inf_{b \in \boldsymbol{m}(\nu)} h(b), \qquad \textit{where} \qquad \boldsymbol{m}(\nu) = \arg\min_{b \in B \cap \boldsymbol{B}_0} \nu(b)$$

Proof. The proof closely follows that of Fang & Santos (2019) lemma S.4.9, found in that paper's supplementary material.

Let $\{t_n\}_{n=1}^\infty \subset \mathbb{R}$ and $\{h_n\}_{n=1}^\infty \subset \ell^\infty(B)$ with $t_n \downarrow 0$ and $\|h_n - h\|_B = \sup_{b \in B} |h_n(b) - h(b)| \to 0$ for some $h \in \mathcal{C}(B)$. By definition, it suffices to show

$$0 = \lim_{n \to \infty} \left| \frac{\iota(\nu + t_n h_n) - \iota(\nu)}{t_n} - \iota'_\nu(h) \right|$$

First notice that

$$\begin{aligned}
&\left| \frac{\iota(\nu + t_n h_n) - \iota(\nu)}{t_n} - \iota'_\nu(h) \right| \\
&\quad = \left| \frac{\inf_{b \in B \cap \mathbf{B}_0} \{\nu(b) + t_n h_n(b)\} - \inf_{b \in B \cap \mathbf{B}_0} \nu(b)}{t_n} - \inf_{b \in \mathbf{m}(\nu)} h(b) \right| \\
&\quad = \left| \frac{\inf_{b \in B \cap \mathbf{B}_0} \{\nu(b) + t_n h_n(b)\} - \inf_{b \in \mathbf{m}(\nu)} \nu(b)}{t_n} - \inf_{b \in \mathbf{m}(\nu)} h(b) \right| \\
&\quad \leq \left| \frac{\inf_{b \in \mathbf{m}(\nu)} \{\nu(b) + t_n h(b)\} - \inf_{b \in \mathbf{m}(\nu)} \nu(b)}{t_n} - \inf_{b \in \mathbf{m}(\nu)} h(b) \right| && (29) \\
&\qquad + \left| \frac{\inf_{b \in B \cap \mathbf{B}_0} \{\nu(b) + t_n h_n(b)\} - \inf_{b \in B \cap \mathbf{B}_0} \{\nu(b) + t_n h(b)\}}{t_n} \right| && (30) \\
&\qquad + \left| \frac{\inf_{b \in B \cap \mathbf{B}_0} \{\nu(b) + t_n h(b)\} - \inf_{b \in \mathbf{m}(\nu)} \{\nu(b) + t_n h(b)\}}{t_n} \right| && (31)
\end{aligned}$$

Consider (29). Notice $\nu(b)$ is flat on $\mathbf{m}(\nu)$, hence $\inf_{b \in \mathbf{m}(\nu)} \{\nu(b) + t_n h(b)\} = \inf_{b \in \mathbf{m}(\nu)} \nu(b) + t_n \inf_{b \in \mathbf{m}(\nu)} h(b)$ and so

$$\frac{\inf_{b \in \mathbf{m}(\nu)} \{\nu(b) + t_n h(b)\} - \inf_{b \in \mathbf{m}(\nu)} \nu(b)}{t_n} = \inf_{b \in \mathbf{m}(\nu)} h(b) \tag{32}$$

it follows that (29) equals zero.

Now consider (30). Apply lemma E.2 to find that

$$\left| \inf_{b \in B \cap \mathbf{B}_0} \{\nu(b) + t_n h_n(b)\} - \inf_{b \in B \cap \mathbf{B}_0} \{\nu(b) + t_n h(b)\} \right| \leq t_n \|h_n - h\|_B$$

divide both sides by t_n to see that

$$\left|\frac{\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh_n(b)\}-\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh(b)\}}{t_n}\right|\le\|h_n-h\|_B\to 0 \tag{33}$$

Lastly consider (31). The extreme value theorem implies $\mathbf{m}(g)=\arg\min_{b\in B\cap\mathbf{B}_0}$ is well defined for any $g\in\mathcal{C}(B)$. Notice that

$$\Psi_\nu:\ell^\infty(B)\to\ell^\infty(B), \qquad\qquad \Psi_\nu(g)=\nu+g$$

is a continuous map, so the Berge maximum theorem (Aliprantis & Border (2006) theorem 17.31) implies the correspondence $g\mapsto\mathbf{m}(\nu+g)$ is compact valued and upper hemicontinuous. Letting $\mathbf{m}(\nu)^\epsilon=\left\{b\in B\cap\mathbf{B}_0\ ;\ \inf_{\tilde{b}\in\mathbf{m}(\nu)}\|b-\tilde{b}\|\le\epsilon\right\}$, upper hemicontinuity and $\|t_nh\|_B=\sup_{b\in B}|t_nh(b)|\to 0$ implies that $\mathbf{m}(\nu+t_nh)\subseteq\mathbf{m}(\nu)^{\delta_n}$ for some sequence $\delta_n\downarrow 0$.[24] Use this to see that

$$\begin{aligned}&\left|\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh(b)\}-\inf_{b\in\mathbf{m}(\nu)}\{\nu(b)+t_nh(b)\}\right|\\ &\qquad\qquad=\inf_{b\in\mathbf{m}(\nu)}\{\nu(b)+t_nh(b)\}-\inf_{b\in\mathbf{m}(\nu)^{\delta_n}}\{\nu(b)+t_nh(b)\}\end{aligned}$$

Let $b_0\in\arg\max_{b\in\mathbf{m}(\nu)}\{\nu(b)+t_nh(b)\}$ be arbitrary. Notice that $\mathbf{m}(\nu+t_nh)\subseteq\mathbf{m}(\nu)^{\delta_n}$ implies

1. there exists $b_1\in\mathbf{m}(\nu+t_nh)$ with $\|b_0-b_1\|\le\delta_n$, and
2. that $\inf_{b\in\mathbf{m}(\nu)^{\delta_n}}\{\nu(b)+t_nh(b)\}=\inf_{b\in\mathbf{m}(\nu+t_nh)}\{\nu(b)+t_nh(b)\}$.

So, $\inf_{b\in\mathbf{m}(\nu)^{\delta_n}}\{\nu(b)+t_nh(b)\}=\inf_{b\in\mathbf{m}(\nu+t_nh),\ \|b_0-b\|\le\delta_n}\{\nu(b)+t_nh(b)\}$.
Let $b_1\in\arg\min_{b\in\mathbf{m}(\nu+t_nh),\ \|b_0-b\|\le\delta_n}\{\nu(b)+t_nh(b)\}$, and notice the display above implies

$$\begin{aligned}&\left|\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh(b)\}-\inf_{b\in\mathbf{m}(\nu)}\{\nu(b)+t_nh(b)\}\right|\\ &\qquad\qquad=\nu(b_0)+t_nh(b_0)-\nu(b_1)-t_nh(b_1)\\ &\qquad\qquad\le t_n\left(h(b_0)-h(b_1)\right)\end{aligned}\tag{34}$$

where the inequality follows because b_0 minimizes $\nu(\cdot)$ over $B\cap\mathbf{B}_0$, and $b_1\in B\cap\mathbf{B}_0$ may not. Divide both sides of (34) by t_n to find

$$\begin{aligned}&\left|\frac{\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh(b)\}-\inf_{b\in\mathbf{m}(\nu)}\{\nu(b)+t_nh(b)\}}{t_n}\right|\\ &\qquad\le h(b_0)-h(b_1)\\ &\qquad\le\sup_{b_0,b_1\in B\cap\mathbf{B}_0,\ \|b_0-b_1\|\le\delta_n}|h(b_0)-h(b_1)|\end{aligned}\tag{35}$$

[24] Recall that $\mathbf{m}$ is compact valued and upper hemicontinuous if and only if for any sequence $\{(\nu+t_nh,b_n)\}_{n=1}^\infty$ with $b_n\in\mathbf{m}(\nu+t_nh)$, $\nu+t_nh\to\nu$, and $b_n\to b$, we have $b\in\mathbf{m}(\nu)$ (see Aliprantis & Border (2006) theorem 17.20 on p. 565). The claim $\mathbf{m}(\nu+t_nh)\subseteq\mathbf{m}(\nu)^{\delta_n}$ for some $\delta_n\downarrow 0$ is equivalent to the claim that there exists a sequence $\delta_n\downarrow 0$ such that for any $\{b_n\}_{n=1}^\infty$ with $b_n\in\mathbf{m}(\nu+t_nh)$, one has $\inf_{\tilde{b}\in\mathbf{m}(\nu)}\|b_n-\tilde{b}\|\le\delta_n$. If it were false, then there would exist a sequence $\{\ddot{b}_n\}_{n=1}^\infty$ with $\ddot{b}_n\in\mathbf{m}(\nu+t_nh)$ such that $\inf_{\tilde{b}\in\mathbf{m}(\nu)}\|\ddot{b}_n-\tilde{b}\|$ does not converge to zero. $\{\ddot{b}_n\}_{n=1}^\infty\subseteq B\cap\mathbf{B}_0$, a compact set, hence there exists a convergent subsequence $\{\ddot{b}_{n_k}\}_{k=1}^\infty$ whose limit is not in $\mathbf{m}(\nu)$. The existence of $\{(\nu+t_{n_k}h,\ddot{b}_{n_k})\}_{k=1}^\infty$ contradicts the fact that $\mathbf{m}$ is upper hemicontinuous and compact valued.

Since h is continuous and defined on a compact set it is uniformly continuous by the Heine-Cantor theorem. Therefore $\sup_{b_0,b_1\in B\cap\mathbf{B}_0,\ \|b_0-b_1\|\le\delta_n}|h(b_0)-h(b_1)|\to 0$.

In summary, (29), (30), (31), (32), (33) and (35) imply

$$\begin{aligned}
&\left|\frac{\iota(\nu+t_nh_n)-\iota(\nu)}{t_n}-\iota'_\nu(h)\right|\\
&\qquad\le\left|\frac{\inf_{b\in\mathbf{m}(\nu)}\{\nu(b)+t_nh(b)\}-\inf_{b\in\mathbf{m}(\nu)}\nu(b)}{t_n}-\inf_{b\in\mathbf{m}(\nu)}h(b)\right|\\
&\qquad\quad+\left|\frac{\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh_n(b)\}-\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh(b)\}}{t_n}\right|\\
&\qquad\quad+\left|\frac{\inf_{b\in B\cap\mathbf{B}_0}\{\nu(b)+t_nh(b)\}-\inf_{b\in\mathbf{m}(\nu)}\{\nu(b)+t_nh(b)\}}{t_n}\right|\\
&\qquad\quad\le 0+\|h_n-h\|_B+\sup_{b_0,b_1\in B\cap\mathbf{B}_0,\ \|b_0-b_1\|\le\delta_n}|h(b_0)-h(b_1)|\to 0
\end{aligned}$$

which completes the proof. □

Lemma E.6 (Uniformity in CLT for processes)**.** *Let $\hat\theta_n,\theta_0:B\to\mathbb{R}^{d_\theta}$, $\hat\theta_n$ random and θ_0 deterministic, with $\hat\theta_n(b),\theta(b)\in\Theta^b$ for each $b\in B$. If*

(i) $\Theta^B=\{(b,\theta)\ ;\ b\in B,\ \theta\in\Theta^b\}$ *is compact,*

(ii) $f(x,b,\theta)$ *is continuous in* (b,θ),

(iii) $\sup_{b\in B}\|\hat\theta_n(b)-\theta_0(b)\|=o_p(1)$, *and*

(iv) $\{f(x,b,\theta)\ ;\ (b,\theta)\in\Theta^B\}$ *is Donsker, has square integrable envelope F, and satisfies $E[f(X,b,\theta)]=0$ for all $(b,\theta)\in\Theta^B$,*

then

$$\sup_{b\in B}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i,b,\hat\theta_n(b))-\frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i,b,\theta_0(b))\right|=o_p(1)$$

Proof. Let $\mathbb{G}_n(b,\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i,b,\theta)$, and $\mathbb{G}$ the tight Gaussian limiting distribution such that $\mathbb{G}_n\xrightarrow{L}\mathbb{G}$ in $\ell^\infty(\Theta^B)$. By assumption, $\hat\theta_n\xrightarrow{p}\theta_0$ in $\ell^\infty(B)$, and hence $(\mathbb{G}_n,\hat\theta_n)\xrightarrow{L}(\mathbb{G},\theta_0)$ in $\ell^\infty(\Theta^B)\times\ell^\infty(B)$ by van der Vaart (2007) theorem 18.10. Define $g:\ell^\infty(\Theta^B)\times\ell^\infty(B)\to\ell^\infty(B)$ pointwise by

$$g(z,\theta)(b)=z(b,\theta(b))-z(b,\theta_0(b)),$$

and notice that for any $z\in\ell^\infty(\Theta^B)$, $g(z,\theta_0)=0\in\ell^\infty(B)$.

Since $(\mathbb{G}_n,\hat\theta_n)\xrightarrow{L}(\mathbb{G},\theta_0)$ in $\ell^\infty(\Theta^B)\times\ell^\infty(B)$, it suffices to show that g is continuous as a map from $\ell^\infty(\Theta^B)\times\ell^\infty(B)$ to $\ell^\infty(B)$ at almost every point in the support of $(\mathbb{G},\theta_0)$. This will imply

$$g(\mathbb{G}_n,\hat\theta_n)\xrightarrow{L}g(\mathbb{G},\theta_0)=0\qquad\qquad\text{in }\ell^\infty(B)$$

by the continuous mapping theorem. Weak convergence to a degenerate limit is equivalent to convergence in probability to that limit (van der Vaart (2007) theorem 18.10), giving the conclusion

$$\sup_{b\in B}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i,b,\hat\theta_n(b))-\frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i,b,\theta_0(b))\right|=\|g(\mathbb{G}_n,\hat\theta_n)\|_B\xrightarrow{p}0$$

For legibility, the argument that $g : \ell^\infty(\Theta^B) \times \ell^\infty(B) \to \ell^\infty(B)$ is continuous at almost every point in the support of $(\mathbb{G}, \theta_0)$ is broken into steps.

1. For g to be continuous at (z, θ), it suffices that $\tilde{z} : \ell^\infty(B) \to \ell^\infty(B)$ defined pointwise by $\tilde{z}(\theta)(b) = z(b, \theta(b))$ is a continuous map at θ.

   Since g is a map between metric spaces, g is continuous at (z, θ) if for any $(z_n, \theta_n) \to (z, \theta)$ in $\ell^\infty(\Theta^B) \times \ell^\infty(B)$,

   $$\|g(z_n, \theta_n) - g(z, \theta)\|_B = \sup_{b \in B} |z_n(b, \theta_n(b)) - z(b, \theta(b))| \to 0$$

   $(z_n, \theta_n) \to (z, \theta)$ in $\ell^\infty(\Theta^B) \times \ell^\infty(B)$ implies that

   $$\sup_{(b,\theta) \in \Theta^B} |z_n(b, \theta) - z(b, \theta)| \to 0, \qquad \text{and} \qquad \sup_{b \in B} \|\theta_n(b) - \theta(b)\| \to 0$$

   and so

   $$\begin{aligned} &\sup_{b \in B} |z_n(b, \theta_n(b)) - z(b, \theta(b))| \\ &\qquad \leq \sup_{b \in B} |z_n(b, \theta_n(b)) - z(b, \theta_n(b))| + \sup_{b \in B} |z(b, \theta_n(b)) - z(b, \theta(b))r| \\ &\qquad \leq \underbrace{\sup_{(b,\theta) \in \Theta^B} |z_n(b, \theta) - z(b, \theta)|}_{\to 0} + \sup_{b \in B} |z(b, \theta_n(b)) - z(b, \theta(b))| \end{aligned}$$

   Therefore g is continuous at any (z, θ) where $\sup_{b \in B} \|\theta_n(b) - \theta(b)\| \to 0$ implies $\sup_{b \in B} |z(b, \theta_n(b)) - z(b, \theta(b))| \to 0$, i.e., where $\tilde{z} : \ell^\infty(B) \to \ell^\infty(B)$ given by $\tilde{z}(\theta)(b) = z(b, \theta(b))$ is a continuous map at θ.

2. For $z \in \ell^\infty(\Theta^B)$ to define a continuous $\tilde{z} : \ell^\infty(B) \to \ell^\infty(B)$ given by $\tilde{z}(\theta)(b) = z(b, \theta(b))$, it suffices that z is uniformly continuous over $(\Theta^B, \|\cdot\|)$.

   This follows from lemma E.3.

3. Almost all sample paths of $\mathbb{G}$ are uniformly continuous in (Θ^B, ρ_2), where

   $$\rho_2((b_1, \theta_1), (b_2, \theta_2)) = E\left[(f(X, b_1, \theta_1) - f(X, b_2, \theta_2))^2\right]^{1/2}$$

   This fact is well known; see van der Vaart & Wellner (1997) section 1.5 (especially addendum 1.5.8 and example 1.5.10) or van der Vaart (2007) lemma 18.15. In other words, for almost all ω, to each $\varepsilon > 0$ there exists $\delta > 0$ such that

   $$\rho_2((b_1, \theta_1), (b_2, \theta_2)) < \delta \implies |\mathbb{G}(b_1, \theta_1, \omega) - \mathbb{G}(b_2, \theta_2, \omega)| < \varepsilon$$

4. $\rho_2 : \Theta^B \times \Theta^B \to \mathbb{R}$ is uniformly continuous.

   First notice that ρ_2 is continuous. Let $((b_1, \theta_1), (b_2, \theta_2)) \in \Theta^B \times \Theta^B$ and $\{((b_{1n}, \theta_{1n}), (b_{2n}, \theta_{2n}))\}_{n=1}^\infty \subset \Theta^B \times \Theta^B$ be such that $((b_{1n}, \theta_{1n}), (b_{2n}, \theta_{2n})) \to ((b_1, \theta_1), (b_2, \theta_2))$. Since $f(x, b, \theta)$ is continuous in (b, θ),

   $$\lim_{n \to \infty} (f(x, b_{1n}, \theta_{1n}) - f(x, b_{2n}, \theta_{2n}))^2 = (f(x, b_1, \theta_1) - f(x, b_2, \theta_2))^2$$

and since $|f(x, b, \theta)| \leq F(x)$ with $E[F(X)] < \infty$,

$$\begin{aligned}
&(f(x, b_{1n}, \theta_{1n}) - f(x, b_{2n}, \theta_{2n}))^2 \\
&\qquad \leq \left| f(x, b_{1n}, \theta_{1n})^2 - 2f(x, b_{1n}, \theta_{1n}) f(x, b_{2n}, \theta_{2n}) + f(x, b_{2n}, \theta_{2n})^2 \right| \\
&\qquad \leq 2F(x)^2 \\
\implies & E\left[(f(X, b_{1n}, \theta_{1n}) - f(X, b_{2n}, \theta_{2n}))^2 \right] \\
&\qquad \leq 4E[F(X)^2] < \infty
\end{aligned}$$

Hence the dominated convergence theorem implies $\lim_{n\to\infty} E\left[(f(X, b_{1n}, \theta_{1n}) - f(X, b_{2n}, \theta_{2n}))^2 \right]^{1/2} = E\left[(f(X, b_1, \theta_1) - f(X, b_2, \theta_2))^2 \right]^{1/2}$, i.e., ρ_2 is continuous at any $((b_1, \theta_1), (b_2, \theta_2)) \in \Theta^B \times \Theta^B$.

ρ_2 is a continuous function defined on the compact set $\Theta^B \times \Theta^B$, and is therefore uniformly continuous by the Heine-Cantor theorem.

5. The preceding two steps imply $\mathbb{G}$ is almost surely uniformly continuous in $(\Theta^B, \|\cdot\|)$.

   Let ω be such that $\mathbb{G}(b, \theta, \omega)$ is uniformly continuous on (Θ^B, ρ_2). The for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\rho_2((b_1, \theta_1), (b_2, \theta_2)) < \delta \implies |\mathbb{G}(b_1, \theta_1, \omega) - \mathbb{G}(b_2, \theta_2, \omega)| < \varepsilon$$

   Use uniform continuity of $\rho_2 : \Theta^B \times \Theta^B \to \mathbb{R}$ to find $\eta > 0$ such that

$$\|(b_1, \theta_1) - (b_2, \theta_2)\| < \eta \implies \rho_2((b_1, \theta_1), (b_2, \theta_2)) < \delta$$

   and notice that $\|(b_1, \theta_1) - (b_2, \theta_2)\| < \eta$ implies $|\mathbb{G}(b_1, \theta_1, \omega) - \mathbb{G}(b_2, \theta_2, \omega)| < \varepsilon$; i.e., $\mathbb{G}(b, \theta, \omega)$ is uniformly continuous on $(\Theta^B, \|\cdot\|)$

In summary, $\mathbb{G}$ is almost surely uniformly continuous on $(\Theta^B, \|\cdot\|)$ and hence the random $\tilde{\mathbb{G}} : \ell^\infty(B) \to \ell^\infty(B)$ defined by $\tilde{\mathbb{G}}(\theta)(b) = \mathbb{G}(b, \theta(b))$ is almost surely a continuous map at θ_0. Therefore g is continuous at almost every point in the support of $(\mathbb{G}, \theta_0)$, implying $g(\mathbb{G}_n, \hat{\theta}_n) \xrightarrow{L} g(\mathbb{G}, \theta_0) = 0$; equivalently,

$$\|g(\mathbb{G}_n, \hat{\theta}_n)\|_B = \sup_{b\in B} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i, b, \hat{\theta}_n(b)) - \frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i, b, \theta_0(b)) \right| = o_p(1)$$

□

Remark E.4. The general proof strategy of lemma E.6 is quite similar to that of lemma 19.24 in van der Vaart (2007).

Lemma E.7 (Sufficient condition for bootstrap confidence interval consistency). *Let $\{X_i\}_{i=1}^n$ be a sequence of random variables, independent of $\{W_i\}_{i=1}^n$. Suppose that*

1. *$\hat{\gamma}_n : \{X_i\}_{i=1}^n \to \mathbb{R}$ satisfies*

$$r_n(\hat{\gamma}_n - \gamma_0) \xrightarrow{L} \mathbb{G}$$

   for some $r_n \uparrow \infty$,

2. *$\hat{\gamma}_n^*: \{X_i, W_i\}_{i=1}^n \to \mathbb{R}$ satisfies*

$$\sup_{f \in BL_1(\mathbb{R})} \left| E\left[f\left(r_n(\hat{\gamma}_n^* - \hat{\gamma}_n)\right) \mid \{X_i\}_{i=1}^n\right] - E\left[f(\mathbb{G})\right]\right| \xrightarrow{p} 0 \tag{36}$$

3. *$c_{1-\alpha} = F^{-1}(1-\alpha)$, where $F^{-1}(p) = \inf\{c\ ;\ F(c) = P(\mathbb{G} \leq c) \geq p\}$, and*

4. *$\hat{c}_{1-\alpha,n} = \hat{F}_n^{-1}(1-\alpha)$, where $\hat{F}_n^{-1}(p) = \inf\left\{c\ ;\ \hat{F}_n(c) = P\left(r_n(\hat{\gamma}_n^* - \hat{\gamma}_n) \leq c\right) \geq p\right\}$*

If $F(c) = P(\mathbb{G} \leq c)$ is continuous and strictly increasing at $c_{1-\alpha}$, then

$$\lim_{n\to\infty} P\left(r_n(\hat{\gamma}_n - \gamma_0) \leq \hat{c}_{1-\alpha,n}\right) = 1 - \alpha$$

Proof. First, suppose that

$$\sup_{f \in \mathrm{BL}_1(\mathbb{R})} \left| E\left[f\left(r_n(\hat{\gamma}_n^* - \hat{\gamma}_n)\right) \mid \{X_i\}_{i=1}^n\right] - E\left[f\left(\mathbb{G}\right)\right]\right| \xrightarrow{a.s.} 0 \tag{37}$$

Let $\mathcal{N} \subset \Omega$ be the negligible set where this convergence fails, and consider $\omega \in \mathcal{N}^c$. Let $\hat{F}_n(c;\omega) = P\left(r_n(\hat{\gamma}_n^* - \hat{\gamma}_n) \leq c \mid \{X_i(\omega)\}_{i=1}^n\right)$, and $\hat{F}_n^{-1}(p;\omega) = \inf\left\{c\ ;\ \hat{F}_n(c;\omega) \geq p\right\}$. Theorem 1.12.4 in van der Vaart & Wellner (1997) shows $Y_n \xrightarrow{L} Y$ if and only if $\sup_{f \in \mathrm{BL}_1} \left|E\left[f(Y_n)\right] - E\left[f(Y)\right]\right| \to 0$, thus $\hat{F}_n(c;\omega) \to F(c)$ for all c where F is continuous. By van der Vaart (2007) lemma 21.2, $\hat{F}_n^{-1}(p;\omega) \to F^{-1}(p)$ at all p where F^{-1} is continuous. F is strictly increasing at c if and only F^{-1} is continuous at p, where $p = F(c)$. Thus $\hat{c}_{1-\alpha,n}(\omega) = \hat{F}_n^{-1}(1-\alpha;\omega) \to F^{-1}(1-\alpha) = c_{1-\alpha}$, i.e. $\hat{c}_{1-\alpha} \xrightarrow{a.s.} c_{1-\alpha}$.

$r_n(\hat{\gamma}_n - \gamma_0) \xrightarrow{L} \mathbb{G}$ and $\hat{c}_{1-\alpha} \xrightarrow{p} c_{1-\alpha}$, so Slutsky's theorem implies $r_n(\hat{\gamma}_n - \gamma_0) - \hat{c}_{1-\alpha} \xrightarrow{L} \mathbb{G} - c_{1-\alpha}$. Since F is continuous at $c_{1-\alpha}$, $P(\mathbb{G} = c_{1-\alpha}) = 0$ and hence the Portmanteau theorem (van der Vaart (2007) theorem 2.2 (vii)) implies $\lim_{n\to\infty} P\left(r_n(\hat{\gamma}_n - \gamma_0) \leq \hat{c}_{1-\alpha,n}\right) = P(\mathbb{G} \leq c_{1-\alpha}) = 1 - \alpha$.

Now, relax (37) back to the assumed (36). Suppose for contradiction that $P\left(r_n(\hat{\gamma}_n - \gamma_0) \leq \hat{c}_{1-\alpha,n}\right) \not\to 1 - \alpha$. Then there exists a subsequence $\{n'\}$ and $\varepsilon > 0$ such that

$$\left|P\left(r_{n'}(\hat{\gamma}_{n'} - \gamma_0) \leq \hat{c}_{1-\alpha,n'}\right) - (1-\alpha)\right| > \varepsilon, \qquad \text{for all } n' \tag{38}$$

But for a further subsequence $\{n''\}$, $\sup_{f \in \mathrm{BL}_1(\mathbb{R})} \left|E\left[f\left(r_{n''}(\hat{\gamma}_{n''}^* - \hat{\gamma}_{n''})\right) \mid \{X_i\}_{i=1}^{n''}\right] - E\left[f\left(\mathbb{G}\right)\right]\right| \xrightarrow{a.s.} 0$ and the argument above implies $P\left(r_{n''}(\hat{\gamma}_{n''} - \gamma_0) \leq \hat{c}_{1-\alpha,n''}\right) \to 1-\alpha$, the desired contradiction. □

Remark E.5. The proof of lemma E.7 is similar to that of lemma 23.3 in van der Vaart (2007).

F Proofs of estimation results

The proofs of the consistency and inference results will make use of the following additional notation.

Recall that, as defined in appendix E, $\ell^\infty(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R}\ ;\ \sup_{x\in\mathcal{X}}|f(x)| < \infty\}$ denotes the set of bounded functions on some set $\mathcal{X}$. $\ell^\infty(\mathcal{X})$ is equipped with the sup-norm: for $f \in \ell^\infty(\mathcal{X})$, $\|f\|_\infty = \|f\|_{\mathcal{X}} = \sup_{x\in\mathcal{X}}|f(x)|$.

The outer minimization in the definition of the breakdown point is the infimum over a restricted set. Define

$$\iota_{B\cap\mathbf{B}_0} : \ell^\infty(B) \to \mathbb{R}, \qquad \iota_{B\cap\mathbf{B}_0}(h) = \inf_{b\in B\cap\mathbf{B}_0} h(b) \tag{39}$$

and notice that under assumption 2, the breakdown point is $\delta^{BP} = \iota_{B\cap\mathbf{B}_0}(\nu)$. The estimator of the breakdown point is $\hat{\delta}^{BP} = \iota_{B\cap\mathbf{B}_0}(\hat{\nu}_n)$.

The population and estimated dual objective are denoted

$$\nu(b,\lambda) = E[\varphi(D, DY, X, b, \lambda, p_D)] \tag{40}$$

$$\hat{\nu}_n(b,\lambda) = \frac{1}{n}\sum_{i=1}^{n} \varphi(D_i, D_iY_i, X_i, b, \lambda, \hat{p}_{D,n}) \tag{41}$$

F.1 Consistency

Theorem 4.1 (Consistency of breakdown point)**.** *Suppose assumptions 1 and 2 hold, and* $\sup_{b\in B}|\hat{\nu}_n(b) - \nu(b)| \xrightarrow{p} 0$. *Then* $\hat{\delta}_n^{BP} \xrightarrow{p} \delta^{BP}$.

Proof. Technical lemma E.2 shows that

$$\iota_{B\cap\mathbf{B}_0} : \ell^\infty(B) \to \mathbb{R}, \qquad\qquad \iota_{B\cap\mathbf{B}_0}(f) = \inf_{b\in B\cap\mathbf{B}_0} f(b)$$

is continuous. Since $\hat{\delta}_n^{BP} = \iota_{B\cap\mathbf{B}_0}(\hat{\nu}_n)$ and $\delta^{BP} = \iota_{B\cap\mathbf{B}_0}(\nu)$ by assumption 2, the result follows by the continuous mapping theorem. □

Lemma F.1 (Unique dual solution)**.** *Suppose assumptions 1 and 2 hold, and for each* $b \in B$, $\{h_j(z,b)\}_{j=1}^{d_g+K}$ *are linearly independent in the sense that for any* $\lambda \in \mathbb{R}^{d_g+K} \setminus \{0\}$,

$$P(\lambda^\intercal h(Z,b) \neq 0 \mid D = 1) > 0 \tag{42}$$

Then for each $b \in B$, $\nu(b,\lambda)$ *is strictly concave in* λ *and hence* $\lambda_0(b) = \arg\max_{\lambda\in\mathbb{R}^{d_g+K}} \nu(b,\lambda)$ *is unique.*

Proof. Let $\lambda_0 \neq \lambda_1$, and $\alpha \in (0,1)$. By strict convexity of f^*,

$$f^*((\alpha\lambda_1 + (1-\alpha)\lambda_0)^\intercal h(z,b)) < \alpha f^*(\lambda_1^\intercal h(z,b)) + (1-\alpha) f^*(\lambda_0^\intercal h(z,b)) \tag{43}$$

for any z where $\lambda_0^\intercal h(z,b) \neq \lambda_1^\intercal h(z,b)$, equivalently, where $(\lambda_0 - \lambda_1)^\intercal h(z,b) \neq 0$. Since $\lambda_0 - \lambda_1 \neq 0$, (42) implies $\{z\ ;\ (\lambda_0 - \lambda_1)^\intercal h(z,b) \neq 0\}$ is a P_1-nonnegligible set. Integrating (43) with respect to P_1 gives

$$E_{P_1}\left[f^*((\alpha\lambda_1 + (1-\alpha)\lambda_0)^\intercal h(Z,b))\right] < \alpha E_{P_1}\left[f^*(\lambda_1^\intercal h(Z,b))\right] + (1-\alpha) E_{P_1}\left[f^*(\lambda_0^\intercal h(Z,b))\right]$$

and hence

$$\begin{aligned}
\nu(b, \alpha\lambda_1 + (1-\alpha)\lambda_0) &= (\alpha\lambda_1 + (1-\alpha)\lambda_0)^\intercal \left(\frac{E_{P_1}[J(D)h(Z,b)]}{1-p_D}\right) \\
&\qquad\qquad - E_{P_1}\left[f^*((\alpha\lambda_1 + (1-\alpha)\lambda_0)^\intercal h(Z,b))\right] \\
&> \alpha\lambda_1^\intercal \left(\frac{E_{P_1}[J(D)h(Z,b)]}{1-p_D}\right) + (1-\alpha)\lambda_0^\intercal \left(\frac{E_{P_1}[J(D)h(Z,b)]}{1-p_D}\right) \\
&\qquad\qquad - \alpha E_{P_1}\left[f^*(\lambda_1^\intercal h(Z,b))\right] - (1-\alpha) E_{P_1}\left[f^*(\lambda_0^\intercal h(Z,b))\right] \\
&= \alpha\nu(b,\lambda_1) + (1-\alpha)\nu(b,\lambda_0)
\end{aligned}$$

$\nu(b,\lambda)$ attains a maximum by 3.1. Any maximizer of a strictly concave function must be unique, completing the proof. □

F.1.1 Consistency with convexity of the value functions

Lemma F.2 (Concavity, pointwise consistency of dual objective)**.** *Let assumption 1 hold. Then for $b \in \boldsymbol{B}$,*

(a) $\hat{\nu}_n(b,\cdot)$ and $\nu(b,\cdot)$ are concave, and

(b) $\hat{\nu}_n(b,\lambda) \xrightarrow{p} \nu(b,\lambda)$ for any λ satisfying $E_{P_1}[|f^*(\lambda^\intercal h(Z,b))|] < \infty$.

Proof. (a) follows from convexity of $f^*(\cdot)$, which implies $\varphi(d,y,x,b,\lambda,p)$ is concave in λ. In detail, let $\lambda_1, \lambda_0 \in \mathbb{R}^{d_g+K}$ be arbitrary and $\alpha \in (0,1)$. Convexity of f^* implies $\varphi(d,y,x,b,\lambda,p)$ is concave in λ:

$$\begin{aligned}
&\varphi(d,y,x,b,\alpha\lambda_1+(1-\alpha)\lambda_0,p)\\
&\qquad= (\alpha\lambda_1+(1-\alpha)\lambda_0)^\intercal\left(\frac{J(d)h(y,x,b)}{1-p}\right) - \frac{d}{p}f^*((\alpha\lambda_1+(1-\alpha)\lambda_0)^\intercal h(y,x,b))\\
&\qquad\geq \alpha\lambda_1^\intercal\left(\frac{J(d)h(y,x,b)}{1-p}\right) + (1-\alpha)\lambda_0^\intercal\left(\frac{J(d)h(y,x,b)}{1-p}\right) - \alpha f^*(\lambda_1^\intercal h(y,x,b)) - (1-\alpha)f^*(\lambda_0^\intercal h(y,x,b))\\
&\qquad= \alpha\varphi(d,y,x,b,\lambda_1,p) + (1-\alpha)\varphi(d,y,x,b,\lambda_0,p)
\end{aligned}$$

It follows that

$$\begin{aligned}
\hat{\nu}_n(b,\alpha\lambda_1+(1-\alpha)\lambda_0) &= \frac{1}{n}\sum_{i=1}^n \varphi(D_i,D_iY_i,X_i,b,\alpha\lambda_1+(1-\alpha)\lambda_0,\hat{p}_{D,n})\\
&\geq \alpha\frac{1}{n}\sum_{i=1}^n \varphi(D_i,D_iY_i,X_i,b,\lambda_1,\hat{p}_{D,n}) + (1-\alpha)\frac{1}{n}\sum_{i=1}^n \varphi(D_i,D_iY_i,X_i,b,\lambda_0,\hat{p}_{D,n})\\
&= \alpha\hat{\nu}_n(b,\lambda_1) + (1-\alpha)\hat{\nu}_n(b,\lambda_0)
\end{aligned}$$

and similarly, $\nu(b,\alpha\lambda_1+(1-\alpha)\lambda_0) \geq \alpha\nu_n(b,\lambda_1) + (1-\alpha)\nu_n(b,\lambda_0)$.

To see *(b)*, first notice that

$$\begin{aligned}
\hat{\nu}_n(b,\lambda) &= \frac{1}{n}\sum_{i=1}^n \varphi(D_i,D_iY_i,X_i,b,\lambda,\hat{p}_{D,n})\\
&= \frac{1}{n}\sum_{i=1}^n \lambda^\intercal\left(\frac{J(D_i)h(D_iY_i,X_i,b)}{1-\hat{p}_{D,n}}\right) - \frac{D_i}{\hat{p}_{D,n}}f^*(\lambda^\intercal h(D_iY_i,X_i,b))\\
&= \frac{\lambda^\intercal}{1-\hat{p}_{D,n}}\left(\frac{1}{n}\sum_{i=1}^n J(D_i)h(D_iY_i,X_i,b)\right) - \frac{1}{\hat{p}_{D,n}}\frac{1}{n}\sum_{i=1}^n D_i f^*(\lambda^\intercal h(D_iY_i,X_i,b))
\end{aligned}$$

The law of large numbers implies $\hat{p}_{D,n} \xrightarrow{p} p_D$,

$$\frac{1}{n}\sum_{i=1}^{n} J(D_i)h(D_iY_i, X_i, b)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} -D_i g(D_iY_i, X_i, b) \\ (1-D_i)\mathbb{1}\{X_i = x_1\} \\ \vdots \\ (1-D_i)\mathbb{1}\{X_i = x_K\} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} -p_D E_{P_1}[g(Y,X,b)] \\ (1-p_D)P(X = x_1 \mid D=0) \\ \vdots \\ (1-p_D)P(X = x_K \mid D=0) \end{pmatrix} = (1-p_D)c(b)$$

and $\frac{1}{n}\sum_{i=1}^{n} D_i f^*(\lambda^\intercal h(D_iY_i, X_i, b)) \xrightarrow{p} p_D E_{P_1}\left[f^*(\lambda^\intercal h(Y,X,b))\right]$. The continuous mapping theorem gives the result:

$$\hat{\nu}_n(b,\lambda) \xrightarrow{p} \lambda^\intercal c(b) - E_{P_1}\left[f^*(\lambda^\intercal h(Z,b))\right] = \nu(b,\lambda)$$

□

Theorem 4.5 (Consistency with convex value function)**.** *Suppose that assumptions 1 and 2 hold, and*

(i) B is a convex, compact subset of int($\boldsymbol{B}$)

(ii) For each $b \in B$, $\{h_j(y,x,b)\}_{j=1}^{d_g+K}$ are linearly independent in the sense that for any $\lambda \in \mathbb{R}^{d_g+K} \setminus \{0\}$,

$$P(\lambda^\intercal h(Y,X,b) \neq 0 \mid D=1) > 0$$

(iii) $\hat{\nu}_n(\cdot)$ and $\nu(\cdot)$ are convex

Then

$$\sup_{b\in B}|\hat{\nu}_n(b) - \nu(b)| \xrightarrow{p} 0$$

Proof. The proof relies on theorem II.1 in appendix II of Andersen & Gill (1982): If $E \subset \mathbb{R}^p$ is open, $\{F_n\}_{n=1}^{\infty}$ is a sequence of random, concave functions mapping E to $\mathbb{R}$ such that for any $x \in E$, $F_n(x) \xrightarrow{p} f(x)$ as $n \to \infty$ for a real function f defined on E, then $f(x)$ is concave and for any compact $A \subset E$, $\sup_{x\in A}|F_n(x) - f(x)| \xrightarrow{p} 0$.

It suffices to show that for $b \in \text{int}(\mathbf{B})$, $\hat{\nu}_n(b) \to \nu(b)$. To see this, notice that $\{\nu_n\}_{n=1}^{\infty}$ and ν being convex implies $\{-\nu_n\}_{n=1}^{\infty}$ and $-\nu$ are concave, and $\hat{\nu}_n(b) \xrightarrow{p} \nu(b)$ implies $\hat{\nu}_n(b) \xrightarrow{p} \nu(b)$. Furthermore,

$$\sup_{b\in B}|\hat{\nu}_n(b) - \nu(b)| = \sup_{b\in B}|-\hat{\nu}_n(b) - (-\nu(b))|$$

$B \subset \text{int}(\mathbf{B})$ is compact, so theorem II.1 of Anderson and Gill (1982) implies that $\sup_{b\in B}|\hat{\nu}_n(b) - \nu(b)| \xrightarrow{p} 0$.

Let $b \in \text{int}(\mathbf{B})$. $\lambda(b)$ is in the interior of $\{\lambda\ ;\ E_{P_1}[|f^*(\lambda^\intercal h(Z,b))|] < \infty\}$ by assumption 2 (ii), which implies that for some $\varepsilon > 0$ the closed ball of radius 2ε centered at $\lambda(b)$, denoted $\bar{B}_{2\varepsilon}(\lambda(b))$, is contained in the interior of $\{\lambda\ ;\ E_{P_1}[|f^*(\lambda^\intercal h(Z,b))|] < \infty\}$. Observe that $\nu(b) = \sup_{\lambda\in\mathbb{R}^{d_g+K}} \nu(b,\lambda) = \sup_{\lambda\in\bar{B}_{2\varepsilon}(\lambda(b))} \nu(b,\lambda)$. Let $\tilde{\lambda}_n(b) \coloneqq \arg\max_{\lambda\in\bar{B}_{2\varepsilon}(\lambda(b))} \hat{\nu}_n(b,\lambda)$ and notice that

(i) $\lambda(b)$ is the unique maximizer of $\nu(b, \lambda)$ by Lemma F.1,

(ii) $\bar{B}_{2\varepsilon}(\lambda(b))$ is compact,

(iii) $\nu(b, \lambda)$ is continuous in λ on $\bar{B}_{2\varepsilon}(\lambda(b))$ (since any concave function is continuous on the interior of its domain)

(iv) Lemma F.2 shows that $\hat{\nu}_n(b, \lambda)$ is concave in λ and $\hat{\nu}_n(b, \lambda) \xrightarrow{p} \nu(b, \lambda)$ on $\{\lambda \ ; \ E_{P_1}[|f^*(\lambda^\intercal h(Z, b))|] < \infty\}$. Theorem II.1 of Andersen & Gill (1982) implies $\sup_{\lambda \in \bar{B}_{2\varepsilon}(\lambda(b))} |\hat{\nu}_n(b, \lambda) - \nu(b, \lambda)| \xrightarrow{p} 0$.

It follows from standard extremum estimator arguments (see, e.g., Newey & McFadden (1994) theorem 2.1) that $\tilde{\lambda}_n(b) \xrightarrow{p} \lambda(b)$.

The last step follows an argument similar to the proof of Theorem 2.7 in Newey & McFadden (1994). With probability approaching one, $\tilde{\lambda}_n(b)$ is in the interior of $\bar{B}_{2\varepsilon}(\lambda(b))$. When this holds, for any $\lambda \notin \bar{B}_{2\varepsilon}(\lambda(b))$, there exists $\alpha \in (0, 1)$ such that $\alpha\tilde{\lambda}_n(b) + (1 - \alpha)\lambda$ is on the boundary of $\bar{B}_{2\varepsilon}(\lambda(b))$, and hence

$$\begin{aligned} &\hat{\nu}_n(b, \tilde{\lambda}_n(b)) \geq \hat{\nu}_n\left(b, \alpha\tilde{\lambda}_n(b) + (1 - \alpha)\lambda\right) \geq \alpha\hat{\nu}_n(b, \tilde{\lambda}_n(b)) + (1 - \alpha)\hat{\nu}_n(b, \lambda) \\ \implies &(1 - \alpha)\hat{\nu}_n(b, \tilde{\lambda}_n(b)) \geq (1 - \alpha)\hat{\nu}_n(b, \lambda) \end{aligned}$$

and hence $\tilde{\lambda}_n(b)$ is the maximand of $\hat{\nu}_n(b, \cdot)$ over the whole space. We then have that, with probability approaching one,

$$\begin{aligned} |\hat{\nu}_n(b) - \nu(b)| &= \left| \sup_{\lambda \in \mathbb{R}^{d_g + K}} \hat{\nu}_n(b, \lambda) - \sup_{\lambda \in \mathbb{R}^{d_g + K}} \nu(b, \lambda) \right| = \left| \sup_{\lambda \in \bar{B}_{2\varepsilon}(\lambda(b))} \hat{\nu}_n(b, \lambda) - \sup_{\lambda \in \bar{B}_{2\varepsilon}(\lambda(b))} \nu(b, \lambda) \right| \\ &\leq \sup_{\lambda \in \bar{B}_{2\varepsilon}(\lambda(b))} |\hat{\nu}_n(b, \lambda) - \nu(b, \lambda)| \xrightarrow{p} 0 \end{aligned}$$

therefore $\hat{\nu}_n(b) \xrightarrow{p} \nu(b)$, which completes the proof. □

F.1.2 Consistency without convexity of the value function

Lemma F.3 (Continuity of lagrange multipliers)**.** *Suppose assumptions 1, 2, and 3 hold. Then* $\lambda(b) = \arg\max_{\lambda \in \mathbb{R}^{d_g + K}} \nu(b, \lambda)$ *is continuous in* b *for all* $b \in B$.

Proof. The proof is by application of the implicit function theorem found in Zeidler (1986) (Theorem 4.B), applied to the first order condition.

Lemma F.1 implies $\lambda(b)$ uniquely maximizes the strictly concave dual objective 40. The moment conditions in assumption 3 (iv) allow application of the dominated convergence theorem (DCT) to find that $\lambda(b)$ uniquely solves the first order condition

$$\begin{aligned} 0 &= \nabla_\lambda \nu(b, \lambda) \\ &= E\left[\nabla_\lambda \varphi(D, DY, X, b, \lambda, p_D)\right] \\ &= \frac{E[J(D)h(DY, X, b)]}{1 - p_D} - \frac{E[D(f^*)'(\lambda^\intercal h(DY, X, b))h(DY, X, b)]}{p_D} \end{aligned}$$

The implicit function theorem found in Zeidler (1986) (theorem 4.B) states that if

(i) The function $\nabla_\lambda \nu(b, \lambda)$ exist on an open neighborhood U of $(b, \lambda(b))$,

(ii) The matrix $\nabla^2_\lambda \nu(b,\lambda)$ exists on U, with $\nabla^2_\lambda \nu(b,\lambda(b))$ invertible, and

(iii) $\nabla_\lambda \nu(b,\lambda)$ and $\nabla^2_\lambda \nu(b,\lambda)$ are continuous at $(b,\lambda(b))$,

then $\lambda(b)$ is the unique solution to (52). If $\nabla_\lambda \nu(b,\lambda)$ is also continuous on U, then $\lambda(b)$ is continuous in a neighborhood of b.

By the DCT again,

$$\begin{aligned}\nabla^2_\lambda \nu(b,\lambda) &= E_{P_1}\left[\nabla^2_\lambda \varphi(D,DY,X,b,\lambda,p_D)\right]\\ &= -\frac{E\left[D(f^*)''(\lambda^\intercal h(DY,X,b))h(DY,X,b)h(DY,X,b)^\intercal\right]}{p_D}\end{aligned}$$

Lemma F.1 shows strict concavity of $\nu(b,\lambda)$ in λ, implying $\nabla^2_\lambda \nu(b,\lambda)$ is negative definite and hence invertible. $(b,\lambda)\mapsto \nabla_\lambda \nu(b,\lambda)$ and $(b,\lambda)\mapsto \nabla^2_\lambda \nu(b,\lambda)$ are continuous on $\bigcup_{b\in B}\operatorname{int}(\bar\Lambda^b)$ by two more applications of the DCT. We conclude by the implicit function theorem that $\lambda(b)$ is continuous in some neighborhood of b. This holds for all $b \in B$, completing the proof. □

Lemma F.4 (Continuous value functions)**.** *Suppose assumption 1, 2, and 3 hold. Then $\nu(b)$ is continuous in b.*

Proof. $(b,\lambda)\mapsto \nu(b,\lambda) = E_{P_1}[\varphi(D,DY,X,b,\lambda,p_D)]$ is continuous by assumption 3 (iv) and the dominated convergence theorem. Lemma F.3 implies $b\mapsto (b,\lambda(b))$ is continuous, so $\nu(b)=\nu(b,\lambda(b))$ is the composition of continuous maps and hence continuous.

□

Lemma F.5 (Uniform consistency of dual objective)**.** *Suppose assumptions 1, 2, and 3 hold. Then*

$$\sup_{(b,\lambda)\in\bar\Lambda^B} |\hat\nu(b,\lambda)-\nu(b,\lambda)| \xrightarrow{p} 0$$

Proof. Recall that

$$\nu(b,\lambda) = E[\varphi(D,DY,X,b,\lambda,p_D)], \qquad \hat\nu_n(b,\lambda) = \frac{1}{n}\sum_{i=1}^n \varphi(D_i,D_iY_i,X_i,b,\lambda,\hat p_{D,n})$$

Apply Technical lemma E.1:

(i) $\{D_i, D_iY_i, X_i\}_{i=1}^n$ are i.i.d. by assumption 1

(ii) $\hat p_{D,n} = \frac{1}{n}\sum_{i=1}^n D_i \xrightarrow{p} E[D_i] = p_D$ by the law of large numbers.

(iii) $a(d,dy,x,p) \coloneqq \sup_{(b,\lambda)\in\bar\Lambda^B}\|\varphi(d,dy,x,b,\lambda,p)-\varphi(d,dy,x,b,\lambda,p_D)\|$ is continuous in p by the maximum theorem.

(iv) The existence of an open $\mathcal{N}$ satisfying $E[\sup_{p\in\mathcal{N}}\sup_{(b,\lambda)\in\bar\Lambda^B}|\varphi(D,DY,X,b,\lambda,p)|] < \infty$ is directly assumed in assumption 3 (iv).

(v) Consider the class of functions $\left\{(d,dy,x)\mapsto \varphi(d,dy,x,b,\lambda,p_D)\ ;\ (b,\lambda)\in\bar\Lambda^B\right\}$. Using assumption 3 (ii) and (iv), we have that $(b,\lambda)\mapsto \varphi(d,dy,x,b,\lambda,p_D)$ is continuous, $\bar\Lambda^B$ is compact, and $(d, dy, x) \mapsto \sup_{(b,\lambda)\in\bar{\Lambda}^B} |\varphi(d, dy, x, b, \lambda, p_D)|$ is an integrable envelope. It follows from Example 19.8 in van der Vaart (2007) that this class is Glivenko-Cantelli, i.e.,

$$\sup_{(b,\lambda)\in\bar{\Lambda}^B} \left\| \frac{1}{n}\sum_{i=1}^{n} \varphi(D_i, D_iY_i, X_i, b, \lambda, p_D) - E[\varphi(D, DY, X, b, \lambda, p_D)] \right\| \xrightarrow{p} 0$$

□

Theorem 4.2 (Consistency). *Suppose assumptions 1, 2, and 3 hold. Then*

$$\sup_{b\in B} \|(\hat{\nu}_n(b), \hat{\lambda}_n(b)) - (\nu(b), \lambda(b))\| \xrightarrow{p} 0$$

Proof. Let $\bar{\lambda}_n(b) = \arg\max_{\lambda\in\bar{\Lambda}^b} \hat{\nu}_n(b, \lambda)$. The proof consists of three steps:

(1) Show $\sup_{b\in B}\|\bar{\lambda}_n(b) - \lambda(b)\| \xrightarrow{p} 0$

(2) Show $\sup_{b\in B}|\hat{\nu}_n(b, \bar{\lambda}_n(b)) - \nu(b, \lambda(b))| \xrightarrow{p} 0$

(3) Show that with probability approaching one, $\sup_{b\in B}\|\hat{\lambda}_n(b) - \bar{\lambda}_n(b)\| = 0$.

For step (1), we will show that for any $\eta > 0$ there exists $\varepsilon > 0$ such that $\sup_{b\in B} \nu(b, \lambda(b)) - \hat{\nu}_n(b, \bar{\lambda}_n(b)) \leq \varepsilon$ implies $\sup_{b\in B}\|\bar{\lambda}_n(b) - \lambda(b)\| < \eta$, and the probability of the former event converges to one. Lemma F.4 that $\nu(b, \lambda(b)) - \nu(b, \lambda)$ is continuous in (b, λ), and

$$\bar{\Lambda}^{B,\eta} := \left\{(b, \lambda) \in \bar{\Lambda}^B,\ \|\lambda - \lambda(b)\| \geq \eta\right\}$$

is compact. It follows by the extreme value theorem that

$$\sup_{(b,\lambda)\in\bar{\Lambda}^{B,\eta}} \nu(b, \lambda(b)) - \nu(b, \lambda)$$

is attained by some (b^l, λ^l). $\lambda(b)$ is the unique maximizer of $\lambda(b, \lambda)$ for each b by Lemma F.1, hence $\varepsilon := \nu(b^l, \lambda(b^l)) - \nu(b^l, \lambda^l) > 0$. It follows that

$$\sup_{b\in B} \nu(b, \lambda(b)) - \nu(b, \bar{\lambda}_n(b)) \leq \varepsilon \qquad \Longrightarrow \qquad \sup_{b\in B}\|\bar{\lambda}_n(b) - \lambda(b)\| < \eta \tag{44}$$

Now notice that

$$\begin{aligned}
\sup_{b\in B} \nu(b, \lambda(b)) - \nu(b, \bar{\lambda}_n(b)) &\leq \sup_{b\in B} \{\nu(b, \lambda(b)) - \hat{\nu}_n(b, \lambda(b))\} + \underbrace{\sup_{b\in B} \left\{\hat{\nu}_n(b, \lambda(b)) - \hat{\nu}_n(b, \bar{\lambda}_n(b))\right\}}_{\leq 0 \text{ by defn of } \bar{\lambda}_n(b)} \\
&\quad + \sup_{b\in B} \left\{\hat{\nu}_n(b, \bar{\lambda}_n(b)) - \nu(b, \bar{\lambda}_n(b))\right\} \\
&\leq \sup_{b\in B} |\hat{\nu}_n(b, \lambda(b)) - \nu(b, \lambda(b))| + \sup_{b\in B} \left|\hat{\nu}_n(b, \bar{\lambda}_n(b)) - \nu(b, \bar{\lambda}_n(b))\right| \\
&\leq 2 \sup_{(b,\lambda)\in\bar{\Lambda}^B} |\hat{\nu}_n(b, \lambda) - \nu(b, \lambda)|
\end{aligned} \tag{45}$$

Lemma F.5 implies that $\sup_{(b,\lambda)\in\bar{\Lambda}^B}|\hat{\nu}_n(b,\lambda)-\nu(b,\lambda)|\leq\frac{\varepsilon}{2}$ with probability approaching one, and on that event, (44) and (45) implies $\sup_{b\in B}\|\bar{\lambda}_n(b)-\lambda(b)\|<\eta$. Therefore $\sup_{b\in B}\|\bar{\lambda}_n(b)-\lambda(b)\|\xrightarrow{p}0$.

Step (2) follows from Lemma F.5, because

$$\begin{aligned}\sup_{b\in B}\left|\hat{\nu}_n(b,\bar{\lambda}_n(b))-\nu(b,\lambda(b))\right| &= \sup_{b\in B}\left|\sup_{\lambda\in\bar{\Lambda}^b}\hat{\nu}_n(b,\lambda)-\sup_{\lambda\in\bar{\Lambda}^b}\nu(b,\lambda)\right|\\ &\leq \sup_{(b,\lambda)\in\bar{\Lambda}^B}|\hat{\nu}_n(b,\lambda)-\nu(b,\lambda)|\xrightarrow{p}0\end{aligned}$$

Finally, step (3) follows an argument similar to the proof of Theorem 2.7 in Newey & McFadden (1994). $\lambda(b)$ is continuous in b by Lemma F.3 and B is compact by assumption 3 (i), so $\lambda(B):=\{\lambda(b)\ ;\ b\in B\}$ is compact. Notice that $\bigcup_{b\in B}\operatorname{int}(\bar{\Lambda}^b)$ is open, hence $\bar{\Lambda}^B\setminus\left(\bigcup_{b\in B}\operatorname{int}(\bar{\Lambda}^b)\right)$ is closed. This set and $\lambda(B)$ have empty intersection, so the distance between these sets is strictly positive, say c. With probability appoaching one, $\sup_{b\in B}\|\bar{\lambda}_n(b)-\lambda(b)\|<c/2$, and on this event $\bar{\lambda}_n(b)\in\operatorname{int}(\bar{\Lambda}^b)$ uniformly over $b\in B$. Since $\hat{\nu}_n(b,\lambda)$ is concave in λ, no λ outside $\operatorname{int}(\bar{\Lambda}^b)$ could make the objective larger than $\bar{\lambda}_n(b)$. This claim holds for all $b\in B$ simultaneously on the event $\sup_{b\in B}\|\bar{\lambda}_n(b)-\lambda(b)\|<c/2$, hence $\hat{\lambda}_n(b)=\bar{\lambda}_n(b)$ for each $b\in B$, or $\sup_{b\in B}\|\hat{\lambda}_n(b)-\bar{\lambda}_n(b)\|=0$. □

F.2 Inference

The proofs in this section make use the following additional notation and results.

Recall that, as defined in appendix E, $\ell^\infty(\mathcal{X})=\{f:\mathcal{X}\to\mathbb{R}\ ;\ \sup_{x\in\mathcal{X}}|f(x)|<\infty\}$ denotes the set of bounded functions on some set $\mathcal{X}$. $\ell^\infty(\mathcal{X})$ is equipped with the sup-norm: for $f\in\ell^\infty(\mathcal{X})$, $\|f\|_\infty=\|f\|_{\mathcal{X}}=\sup_{x\in\mathcal{X}}|f(x)|$.

The subspace of bounded functions taking values in $\mathbb{R}^K$ for some $K\in\mathbb{N}$ is the product space $\underbrace{\ell^\infty(\mathcal{X})\times\ldots\times\ell^\infty(\mathcal{X})}_{K\text{ times}}$, but can also be viewed as a process on $\ell^\infty(\mathcal{X}\times\{1,\ldots,K\})$. The latter notation makes it clear that standard empirical process results, typically stated for scalar-valued processes, apply.

Assumption 3 is maintained for all inference results. Since B is compact, the set of continuous functions on B are also bounded by the extreme value theorem and hence forms of a subpace of $\ell^\infty(B)$. This subspace is denoted

$$\mathcal{C}(B)=\{f:B\to\mathbb{R}\ ;\ f\text{ is continuous}\}$$

For a $K\times K$ real matrix A, let $\alpha_1(A)\geq\ldots\geq\alpha_K(A)$ be the ordered eigenvalues of A. Let $\sigma_1(A)\geq\ldots\geq\sigma_K(A)\geq0$ be the ordered singular values of A, given by $\sigma_k=\sqrt{\alpha_k(A^\intercal A)}$. Let $\|A\|_o=\sup_{x\ ;\ \|x\|_2=1}\|Ax\|_2$ be the operator norm of A, and $\|A\|_{\max}=\max_{ij}|a_{ij}|$, where $a_{ij}\in\mathbb{R}$ is the entry in the i-th row and j-th column of A.

Recall that all norms on finite dimensional vector spaces are strongly equivalent, implying that if $\|\cdot\|_1$ and $\|\cdot\|_2$ are any norms on $\mathbb{R}^{K\times K}$, there exist constants $c,C>0$ such that $c\|A\|_1\leq\|A\|_2\leq C\|A\|_1$ for any matrix $A\in\mathbb{R}^{K\times K}$. If $A:T\to\mathbb{R}^{K\times K}$ for some set T, it follows that $E[\sup_t\|A(t)\|]<\infty$ for any norm if and only if $E[\sup_t\|A(t)\|_{\max}]<\infty$. Notice that strong equivalence with $\|\cdot\|_{\max}$ implies that, for any submatrix $\tilde{A}(t)$ of $A(t)$, $E[\sup_t\|A(t)\|]<\infty$ implies $E[\sup_t\|\tilde{A}(t)\|]<\infty$.

Recall that $\|A\|_o=\sigma_1(A)$, and that for invertible A and any $k=1,\ldots,K$, $\frac{1}{\sigma_k(A)}$ is a singular value of A^{-1}. These imply $\|A^{-1}\|_o=\frac{1}{\sigma_K(A)}$.

Lemma F.6 (Bounds on matrix norms)**.** *Suppose assumptions 1, 2, 3, and 4 hold. Then* $\sup_{b\in B}\|\Phi(b)\|_o < \infty$ *and* $\sup_{b\in B}\|\Phi(b)^{-1}\|_o < \infty$

Proof. Recall that,

$$\Phi(b) = E\left[\nabla_\theta \phi(D, DY, X, b, \theta)\right]$$

To see that $\sup_{b\in B}\|\Phi(b)\|_o < \infty$, use convexity of norms and Jensen's inequality to see that

$$\begin{aligned}\sup_{b\in B}\|E\left[\nabla_\theta \phi(D, DY, X, b, \theta(b))\right]\| &\leq E\left[\sup_{b\in B}\|\nabla_\theta \phi(D, DY, X, b, \theta(b))\|\right] \\ &\leq E\left[\sup_{b\in B}\sup_{\theta\in\Theta^b}\|\nabla_\theta \phi(D, DY, X, b, \theta)\|\right]\end{aligned}$$

$E\left[\sup_{b\in B}\sup_{\theta\in\Theta^b}\|\nabla_\theta \phi(D, DY, X, b, \theta)\|\right] < \infty$ is implied by Assumption 4 (v) and Jensen's inequality.

Establishing $\sup_{b\in B}\|\Phi(b)^{-1}\|_o < \infty$ is more involved. First, note that

$$\begin{aligned}\nabla_\theta \phi(d, dy, x, b, \theta) &= \left[\nabla_{(\nu,\lambda,p)}\begin{pmatrix}\varphi(d, dy, x, b, \lambda, p) - \nu \\ \nabla_\lambda \varphi(d, dy, x, b, \lambda, p) \\ d - p\end{pmatrix}\right] \\ &= \begin{bmatrix}-1 & \nabla_\lambda \varphi(d, dy, x, b, \lambda, p)^\intercal & \nabla_p \varphi(d, dy, x, b, \lambda, p) \\ 0 & \nabla^2_\lambda \varphi(d, dy, x, b, \lambda, p) & \nabla_p \nabla_\lambda \varphi(d, dy, x, b, \lambda, p) \\ 0 & 0 & -1\end{bmatrix} \qquad (46)\end{aligned}$$

Therefore

$$\begin{aligned}\Phi(b) &= E\left[\nabla_\theta \phi(D, DY, X, b, \lambda, p_D)\right] \\ &= \begin{bmatrix}-1 & 0 & E\left[\nabla_p \varphi(D, DY, X, b, \lambda(b), p_D)\right] \\ 0 & E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right] & E\left[\nabla_p \nabla_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right] \\ 0 & 0 & -1\end{bmatrix}\end{aligned}$$

where $E\left[\nabla_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)^\intercal\right] = 0$ is the first order condition of the dual problem.

Consider the middle matrix: $E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]$, and notice that the dominated convergence theorem and assumption 3 (iv) implies this equals $\nabla^2_\lambda \nu(b, \lambda(b))$. Lemma F.3 shows $\lambda(b)$ is continuous in b, hence $b \mapsto \nabla^2_\lambda \nu(b, \lambda(b))$ is continuous. The mapping from matrices to eigenvalues is continuous (see Bhatia (1997) Corollary III.2.6, used in Lemma E.4) and B is compact, so the extreme value theorem implies $\sup_{b\in B}\alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right)$ is attained by some b^s. Lemma F.1 showed $\nu(b, \lambda)$ is strictly concave in λ, hence

$$\sup_{b\in B}\alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right) = \alpha_1\left(\nabla^2_\lambda \nu(b^s, \lambda(b^s))\right) < 0$$

which implies $\nabla^2_\lambda \nu(b, \lambda(b)) = E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]$ is invertible for each $b \in B$.

With this invertability claim, it's straightforward to verify that $\Phi(b)^{-1}$ is given by

$$\Phi(b)^{-1} =$$

$$\begin{bmatrix}-1 & 0 & -E\left[\nabla_p \varphi(D, DY, X, b, \lambda(b), p_D)\right] \\ 0 & E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]^{-1} & E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]^{-1} E\left[\nabla_p \nabla_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right] \\ 0 & 0 & -1\end{bmatrix}$$

Recall that for any conformable matrices,

$$\left\| \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right\| \leq \|A_{11}\|_o + \|A_{12}\|_o + \|A_{21}\|_o + \|A_{22}\|_o$$
$$\|AB\|_o \leq \|A\|_o \|B\|_o$$

and so

$$\begin{aligned} &\sup_{b\in B} \|\Phi(b)^{-1}\| \\ &\quad \leq 2 + \sup_{b\in B} \Big| E\left[\nabla_p \varphi(D, DY, X, b, \lambda(b), p_D)\right] \Big| + \sup_{b\in B} \left\| E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]^{-1} \right\| \\ &\qquad + \sup_{b\in B} \left\| E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]^{-1} \right\| \sup_{b\in B} \| E\left[\nabla_p \nabla_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right] \| \end{aligned} \tag{47}$$

$\sup_{b\in B} \|\Phi(b)\|_o < \infty$, argued above, implies $\sup_{b\in B} |E\left[\nabla_p \varphi(D, DY, X, b, \lambda(b), p_D)\right]| < \infty$ and $\sup_{b\in B} \|E\left[\nabla_p \nabla_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]\| < \infty$. Moreover, since $E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda(b), p_D)\right]^{-1} = \left(\nabla^2_\lambda \nu(b, \lambda(b))\right)^{-1}$ is symmetric and negative definite,

$$\begin{aligned} \sup_{b\in B} \left\| \left(\nabla^2_\lambda \nu(b, \lambda(b))\right)^{-1} \right\|_o &= \sup_{b\in B} \frac{1}{|\alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right)|} = \frac{1}{\inf_{b\in B} |\alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right)|} \\ &= \frac{1}{|\sup_{b\in B} \alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right)|} \end{aligned}$$

which is finite, since $\sup_{b\in B} \alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right) < 0$ as argued above. Therefore (47) shows that $\sup_{b\in B} \|\Phi(b)^{-1}\| < \infty$. □

Lemma F.7 (Uniform consistency of jacobian terms)**.** *Suppose assumption 1, 2, 3, and 4 hold. Then*

$$\sup_{b\in B} \sup_{\theta\in\Theta^b} \left\| \frac{1}{n} \sum_{i=1}^n \nabla_\theta \phi(D_i, D_i Y_i, X_i, b, \theta) - E\left[\nabla_\theta \phi(D, DY, X, b, \theta)\right] \right\| \xrightarrow{p} 0$$

Proof. By verifying the conditions of example 19.8 in van der Vaart (2007).

Notice that $\Theta^B = \left\{(b, \theta) \,;\, b \in B, \theta \in \Theta^b\right\} = \left\{(b, \nu, \lambda, p) \,;\, (b, \lambda) \in \bar{\Lambda}^B, \nu \in [0, \mathcal{V}], p \in [\underline{p}, \overline{p}]\right\}$ is compact. It's straightforward to verify that $(b, \theta) \mapsto \nabla_\theta \phi(d, dy, x, b, \theta)$ is continuous for any $(b, \theta) \in \Theta^B$, by examination of (46) and

$$\begin{aligned} \nabla_p \varphi(d, dy, x, b, \lambda, p) &= \frac{\lambda^\intercal J(d) h(dy, x, b)}{(1-p)^2} + \frac{d}{p^2} f^*(\lambda^\intercal h(dy, x, b)) \\ \nabla_\lambda \varphi(d, dy, x, b, \lambda, p) &= \frac{J(d) h(dy, x, b)}{1-p} - \frac{d}{p} (f^*)'\left(\lambda^\intercal h(dy, x, b)\right) h(dy, x, b) \\ \nabla^2_\lambda \varphi(d, dy, x, b, \lambda, p) &= -\frac{d}{p} (f^*)''\left(\lambda^\intercal h(dy, x, b)\right) h(dy, x, b) h(dy, x, b)^\intercal \\ \nabla_p \nabla_\lambda \varphi(d, dy, x, b, \lambda, p) &= \frac{J(d) h(dy, x, b)}{(1-p)^2} + \frac{d}{p^2} (f^*)'\left(\lambda^\intercal h(dy, x, b)\right) h(dy, x, b) \end{aligned}$$

Finally, assumption 4 (v) implies $E\left[\sup_{(b,\theta)\in\Theta^B} \left\|\nabla_{(b,\theta)} \phi(D, DY, X, b, \theta)\right\|\right] < \infty$ by Jensen's inequality, which in turn implies $E\left[\sup_{(b,\theta)\in\Theta^B} \|\nabla_\theta \phi(D, DY, X, b, \theta)\|\right] < \infty$ by the reasoning given at the start of Appendix F.2. Thus $\left\{\nabla_\theta \phi(D_i, D_iY_i, X_i, b, \theta) \; ; \; (b,\theta) \in \Theta^b\right\}$ is a special case of example 19.8 in van der Vaart (2007), which gives the result. □

Lemma F.8 (Donsker influence functions)**.** *Suppose assumptions 1, 2, 3, and 4 hold. Then the class of functions*

$$\left\{\phi(D, DY, X, b, \theta) \; ; \; b \in B, \theta \in \Theta^b\right\}$$

is Donsker.

Proof. By verifying the conditions of van der Vaart (2007) example 19.7.

Θ^B is a compact subset of a finite dimensional space, hence bounded. Let $(b_1,\theta_1),(b_2,\theta_2) \in \Theta^B$ and apply the mean value theorem (e.g., Coleman (2012) Corollary 3.2) to find

$$\begin{aligned}
&\|\phi(d, dy, x, b_1, \theta_1) - \phi(d, dy, x, b_2, \theta_2)\| \\
&\quad \leq \left[\sup_{t\in(0,1)} \left\|\nabla_{(b,\theta)}\phi(d, dy, x, b_0 + t(b_1 - b_0), \theta_0 + t(\theta_1 - \theta_0))\right\|_o\right] \|(b_1,\theta_1) - (b_2,\theta_2)\| \\
&\quad \leq \left[\sup_{b\in B}\sup_{\theta\in\Theta^b} \left\|\nabla_{(b,\theta)}\phi(d, dy, x, b, \theta)\right\|_o\right] \|(b_1,\theta_1) - (b_2,\theta_2)\|
\end{aligned}$$

$E\left[\sup_{b\in B}\sup_{\theta\in\Theta^b} \left\|\nabla_{(b,\theta)}\phi(D, DY, X, b, \theta)\right\|_o\right] < \infty$ by assumption 4 (v) and Jensen's inequality. Therefore $\left\{\phi(D, DY, X, b, \theta) \; ; \; b \in B, \theta \in \Theta^b\right\}$ is a special case of van der Vaart (2007) example 19.7. □

F.2.1 Asymptotic Distribution

Theorem 4.6 (Asymptotic distribution of the value function)**.** *Suppose assumptions 1, 2, 3, and 4 hold. Then as a process in $\ell^\infty(B \times \mathcal{I})$,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathbb{G}$$

where $\mathbb{G}$ is a mean zero tight Gaussian process, with covariance function

$$\begin{aligned}
&Cov(\mathbb{G}(b_1, i_1), \mathbb{G}(b_2, i_2)) \\
&\quad = E\left[\left(\Phi(b_1)^{-1}\right)^{(i_1)} \phi\left(D, DY, X, b_1, \theta(b_1)\right) \left[\left(\Phi(b_2)^{-1}\right)^{(i_2)} \phi\left(D, DY, X, b_2, \theta(b_2)\right)\right]\right]
\end{aligned}$$

where $\left(\Phi(b)^{-1}\right)^{(i)}$ is the i-th row of the matrix $\Phi(b)^{-1} = E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0)\right]^{-1}$.

Proof. For legibility, the proof is divided into six steps:

1. Mean value theorem.

   For each $b \in B$, apply the mean value theorem to each coordinate of $0 = \frac{1}{n}\sum_{i=1}^n \phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b))$

and stack the results to obtain

$$0 = \frac{1}{n}\sum_{i=1}^{n} \phi(D_i, D_iY_i, X_i, b, \theta(b)) \\ + \underbrace{\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \phi^{(1)}(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(1)}(b)) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \phi^{(d_g+K+2)}(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(d_g+K+2)}(b)) \end{bmatrix}}_{:=\bar{\Phi}_n(b)} \left(\hat{\theta}_n(b) - \theta(b)\right) \tag{48}$$

where $\bar{\theta}_n^{(k)}(b) = \theta_0(b) + a_n^{(k)}(b)(\hat{\theta}_n(b) - \theta_0(b))$ for some measurable $a_n^{(k)}(b) \in (0,1)$.[25] Since $\bar{\Lambda}^b$ is convex, so is Θ^b, and therefore $\bar{\theta}_n^k(b) \in \Theta^b$ for each b. The triangle inequality implies that for any k,

$$\begin{aligned} \|\bar{\theta}_n^{(k)}(b) - \theta_0(b)\| &\leq a_n^{(k)}(b)\|\hat{\theta}_n(b) - \theta_0(b)\| + (1 - a_n^{(k)}(b))\|\hat{\theta}_n(b) - \theta_0(b)\| \\ &\leq 2\|\hat{\theta}_n(b) - \theta_0(b)\| \xrightarrow{p} 0 \end{aligned}$$

where $\|\hat{\theta}_n(b) - \theta_0(b)\| \xrightarrow{p} 0$ follows from theorem 4.2.

2. Show $\sup_{b\in B} \left\|\bar{\Phi}_n(b) - \Phi(b)\right\|_o \xrightarrow{p} 0$.

   First notice that

$$\bar{\Phi}_n(b) = \sum_{k=1}^{d_g+K+2} e_{kk}\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \phi(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(k)}(b)) \tag{49}$$

where e_{kk} is the square $K \times K$ matrix whose (k,k)-th entry is one and all other entries are zero.[26]

[25] See Newey & McFadden (1994) footnote 25.

[26] When premultiplying a matrix A, e_{kk} "selects" the k-th row. For example,

$$e_{22}A = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1K} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2K} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3K} \\ \vdots & & & \ddots & \vdots \\ a_{K1} & a_{K2} & a_{K3} & \dots & a_{KK} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2K} \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

For any k,

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\phi(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(k)}(b)) - E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0(b))\right]\right\|_o \\
&\qquad\leq \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\phi(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(k)}(b)) - E\left[\nabla_\theta\phi(D, DY, X, b, \bar{\theta}_n^{(k)}(b))\right]\right\|_o \\
&\qquad\qquad + \left\|E\left[\nabla_\theta\phi(D, DY, X, b, \bar{\theta}_n^{(k)}(b))\right] - E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0(b))\right]\right\|_o
\end{aligned}$$

$$\leq \sup_{(b,\theta)\in\Theta^B}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\phi(D_i, D_iY_i, X_i, b, \theta) - E\left[\nabla_\theta\phi(D, DY, X, b, \theta)\right]\right\|_o \tag{50}$$

$$+ \sup_{b\in B}\left\|E\left[\nabla_\theta\phi(D, DY, X, b, \bar{\theta}_n^{(k)}(b))\right] - E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0(b))\right]\right\|_o \tag{51}$$

Lemma F.7 shows the term in (50) converges in probability to zero. Let $\psi : \Theta^B \to \mathbb{R}$ be given by

$$\psi(b, \theta) = \left\|E\left[\nabla_\theta\phi(D, DY, X, b, \theta)\right] - E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0(b))\right]\right\|_o$$

Let $\tilde{\psi} : \ell^\infty_{\Theta^B}(B)^{d_g+K+2} \to \ell^\infty(B)$ defined pointwise by $\tilde{\psi}(\theta)(b) = \psi(\theta(b))$. Notice that ψ is continuous by the DCT and assumption 4 (v), and since Θ^B is compact ψ is in fact uniformly continuous by the Heine-Cantor theorem. By lemma E.3, this suffices for $\tilde{\psi}$ to be continuous continuous. Since $\|\bar{\theta}_n^{(k)} - \theta_0\|_B = \sup_{b\in B}\|\bar{\theta}_n^{(k)}(b) - \theta_0(b)\| \xrightarrow{p} 0$, the continuous mapping theorem then implies $\tilde{\psi}(\bar{\theta}_n^{(k)}) \xrightarrow{p} \tilde{\psi}(\theta_0) = 0 \in \ell^\infty(B)$. The term in (51) equals $\|\tilde{\psi}(\bar{\theta}_n^{(k)})\|_B$, which thus also converges in probability to zero.

Use this to see that

$$\begin{aligned}
&\left\|\bar{\Phi}_n(b) - \Phi(b)\right\|_o \\
&\quad= \left\|\sum_{k=1}^{d_g+K+2} e_{kk}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\phi(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(k)}(b)) - E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0(b))\right]\right)\right\|_o \\
&\quad\leq \sum_{k=1}^{d_g+K+2}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\phi(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(k)}(b)) - E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0(b))\right]\right\|_o \\
&\quad\leq (d_g + K + 2)\max_k\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\phi(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{(k)}(b)) - E\left[\nabla_\theta\phi(D, DY, X, b, \theta_0(b))\right]\right\|_o \\
&\quad\xrightarrow{p} 0
\end{aligned}$$

3. Uniform linearization.

   Lemma F.6 shows $\sup_{b\in B}\|\Phi(b)\|_o < \infty$ and $\sup_{b\in B}\left\|\Phi(b)^{-1}\right\|_o < \infty$. With $\sup_{b\in B}\left\|\bar{\Phi}_n(b) - \Phi(b)\right\|_o \xrightarrow{p} 0$ shown in the previous step, lemma E.4 implies that with probability approaching one,

$\bar{\Phi}_n(b)^{-1}$ is well defined as a function on B. When this holds, rearrange (48) to find

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n(b) - \theta(b)) &= \bar{\Phi}_n(b)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \theta(b)) \\ &= G_n(b) + R_n(b) \\ \text{where } G_n(b) &= \Phi(b)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \theta(b)), \\ \text{and } R_n(b) &= \left[\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \theta(b))\end{aligned}$$

holds uniformly over $b \in B$.

4. The influence functions are Donsker with the claimed limit.

   Define $\tilde{G}_n$ pointwise as $\tilde{G}_n(b,i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi^{(i)}(D_i, D_iY_i, X_i, b, \theta(b))$. $\{\phi(d, dy, x, b, \theta(b))\ ;\ b \in B\}$ is a subset of the class considered in lemma F.8, and is hence Donsker (see van der Vaart & Wellner (1997) theorem 2.10.1). Thus, $\tilde{G}_n \xrightarrow{L} \tilde{\mathbb{G}}$ in $\ell^\infty(B \times \mathcal{I})$, where $\tilde{\mathbb{G}}$ is a mean zero tight Gaussian process with covariance function

   $$\text{Cov}(\tilde{\mathbb{G}}(b_1, i_1), \tilde{\mathbb{G}}(b_2, i_2)) = E\left[\phi^{(i_1)}(D, DY, X, b_1, \theta(b_1))\phi^{(i_2)}(D, DY, X, b_2, \theta(b_2))\right]$$

   Now define

   $$L : \ell^\infty(B \times \mathcal{I}) \to \ell^\infty(B \times \mathcal{I}), \qquad L(H)(b) = \Phi(b)^{-1}H(b)$$

   Note that L is a linear operator. Lemma F.6 shows $\sup_{b\in B}\|\Phi(b)^{-1}\|_o < \infty$, which along with

   $$\|LH\|_B = \sup_{b\in B}\left\|\Phi(b)^{-1}H(b)\right\|_o \le \sup_{b\in B}\left\|\Phi(b)^{-1}\right\|_o \sup_{b\in B}\|H(b)\|_o = \left(\sup_{b\in B}\|\Phi(b)^{-1}\|_o\right)\|H\|_B$$

   shows that L is bounded and hence continuous. The continuous mapping theorem then implies

   $$L(\tilde{G}_n) \xrightarrow{L} L(\tilde{\mathbb{G}})$$

   where $L(\tilde{\mathbb{G}})$ is a mean zero tight Gaussian process on $\ell^\infty(B \times \mathcal{I})$. Letting $(\Phi(b))^{(i)}$ be the i-th row of the matrix $\Phi(b)^{-1}$, the covariance function of $L(\tilde{\mathbb{G}})$ is

   $$\begin{aligned}&\text{Cov}\left(L(\tilde{\mathbb{G}})(b_1, i_1), L(\tilde{\mathbb{G}})(b_2, i_2)\right) \\ &\quad = E\left[\left(\Phi(b_1)^{-1}\right)^{(i_1)}\phi(D, DY, X, b_1, \theta(b_1))\left[\left(\Phi(b_2)^{-1}\right)^{(i_2)}\phi(D, DY, X, b_2, \theta(b_2))\right]\right]\end{aligned}$$

   Notice that the marginals of $L(\tilde{\mathbb{G}})$ are equal in distribution to those of $\mathbb{G}$ in the theorem statement. By van der Vaart & Wellner (1997) Lemma 1.5.3, this implies the two distributions are the same and hence $G_n \xrightarrow{L} \mathbb{G}$.

5. Uniform linearization remainder control.

   Since $\{\phi(d, dy, x, b, \theta(b))\ ;\ b \in B\}$ is Donsker, $\sup_{b\in B}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \theta(b))\right\| = O_p(1)$ by the continuous mapping theorem. Lemma F.7 implies $\sup_{b\in B}\left\|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right\| = o_p(1)$, thus

$$\begin{aligned}\sup_{b\in B}\|R_n(b)\| &= \sup_{b\in B}\left\|\left[\bar{\Phi}_n(b)^{-1}-\Phi(b)^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i,D_iY_i,X_i,b,\theta(b))\right\| \\ &\leq \sup_{b\in B}\left\|\bar{\Phi}_n(b)^{-1}-\Phi(b)^{-1}\right\|\sup_{b\in B}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i,D_iY_i,X_i,b,\theta(b))\right\| \\ &\xrightarrow{p} 0\end{aligned}$$

i.e., $R_n \xrightarrow{p} 0$ as an element of $\ell^\infty(B\times\mathcal{I})$.

6. Conclusion.

   As elements of $\ell^\infty(B\times\mathcal{I})$, $G_n \xrightarrow{L} \mathbb{G}$ and $R_n \xrightarrow{p} 0$, so

   $$(G_n, R_n) \xrightarrow{L} (\mathbb{G}, 0)$$

   by van der Vaart (2007) theorem 18.10. The continuous mapping thoerem (van der Vaart (2007) theorem 18.11) then implies

   $$\sqrt{n}(\hat{\theta}_n-\theta) = G_n + R_n \xrightarrow{L} \mathbb{G}+0$$

   which concludes the proof.

□

Lemma F.9 (Support of value function asymptotic distribution)**.** *Suppose assumption 1, 2, 3, and 4 hold. Let $\mathbb{G}_\nu$ be the mean zero Gaussian process on B defined by $\mathbb{G}_\nu(b)=\mathbb{G}(b,1)$. Then $\sqrt{n}(\hat{\nu}_n-\nu)\xrightarrow{L}\mathbb{G}_\nu$, where $P(\mathbb{G}_\nu\in\mathcal{C}(B))=1$.*

Proof. Theorem 4.6 and the continuous mapping theorem implies $\sqrt{n}(\hat{\nu}_n-\nu)\xrightarrow{L}\mathbb{G}_\nu$. The Portmanteau theorem (van der Vaart & Wellner (1997) theorem 1.3.4) shows that $\sqrt{n}(\hat{\nu}_n-\nu)\xrightarrow{L}\mathbb{G}_\nu$ if and only if for all closed sets F,

$$\limsup_{n\to\infty} P^*\left(\sqrt{n}(\hat{\nu}_n-\nu)\in F\right)\leq P\left(\mathbb{G}_\nu\in F\right).$$

$\mathcal{C}(B)$ is a closed subset of $\ell^\infty(B)$. Lemma F.4 shows ν is continuous, so it suffices to show that $\hat{\nu}_n$ is continuous with probability approaching one.

This follows from the implicit function theorem applied to the first order conditions. Similarly to the proof of lemma F.3, $\hat{\lambda}_n(b)$ must solve the first order condition for each $b\in B$:

$$\begin{aligned}0 &= \nabla_\lambda\hat{\nu}_n(b,\hat{\lambda}_n(b)) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{J(D_i)h(D_iY_i,X_i,b)}{1-p_{\hat{D},n}}-\frac{D_i}{\hat{p}_{D,n}}(f^*)'\left(\hat{\lambda}_n(b)^\intercal h(D_iY_i,X_i,b)\right)h(D_iY_i,X_i,b)\end{aligned} \tag{52}$$

The implicit function theorem found in Zeidler (1986) theorem 4.B states that if

(i) The function $\nabla_\lambda\hat{\nu}_n(b,\lambda)$ exists on an open neighborhood U of $(b,\hat{\lambda}_n(b))$,

(ii) The matrix $\nabla^2_\lambda\hat{\nu}_n(b,\lambda)$ exists on U, with $\nabla^2_\lambda\hat{\nu}_n(b,\hat{\lambda}_n(b))$ invertible, and

(iii) $\nabla_\lambda \hat{\nu}_n(b,\lambda)$ and $\nabla^2_\lambda \hat{\nu}_n(b,\lambda)$ are continuous at $(b, \hat{\lambda}_n(b))$,

then $\hat{\lambda}_n(b)$ is the unique solution to (52). If $\nabla_\lambda \hat{\nu}_n(b,\lambda)$ is also continuous on U, then $\hat{\lambda}_n(b)$ is continuous in a neighborhood of b.

For any $b \in B$ and any n, it is clear that (i) and (iii) hold, that $\nabla^2_\lambda \hat{\nu}_n(b,\lambda)$ exists, and that $(b,\lambda) \mapsto \nabla_\lambda \hat{\nu}_n(b,\lambda)$ is continuous. It remains to show that, with probability approaching one, $\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b))$ is invertible for all $b \in B$ simulataneously. For legibility, this argument is broken into steps.

1. $\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b))$ is uniformly consistent for $\nabla^2_\lambda \nu(b, \lambda(b))$.

   First notice

$$\begin{aligned}\sup_b \|\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b)) - \nabla^2_\lambda \nu(b, \lambda(b))\| &\leq \sup_b \|\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b)) - \nabla^2_\lambda \nu(b, \hat{\lambda}_n(b))\| \\ &\qquad + \sup_b \|\nabla^2_\lambda \nu(b, \hat{\lambda}_n(b)) - \nabla^2_\lambda \nu(b, \lambda(b))\| \\ &\leq \sup_{(b,\lambda)\in\bar{\Lambda}^B} \|\nabla^2_\lambda \hat{\nu}_n(b,\lambda) - \nabla^2_\lambda \nu(b,\lambda)\| \\ &\qquad + \sup_b \|\nabla^2_\lambda \nu(b, \hat{\lambda}_n(b)) - \nabla^2_\lambda \nu(b, \lambda(b))\|\end{aligned}$$

   Lemma F.7 implies that $\sup_{(b,\lambda)\in\bar{\Lambda}^B} \|\nabla^2_\lambda \hat{\nu}_n(b,\lambda) - \nabla^2_\lambda \nu(b,\lambda)\| \xrightarrow{p} 0$. For the second term, notice that $\psi(b,t) = \|\nabla^2_\lambda \nu(b,t) - \nabla^2_\lambda \nu(b,\lambda(b))\|$ is continuous by assumption 3 (iv), the DCT, and lemma F.3. The Heine-Cantor theorem then implies ψ is uniformly continuous on the compact $\bar{\Lambda}^B$. Thus

$$\tilde{\psi} : \ell^\infty_{\bar{\Lambda}^B}(B)^{d_g+K} \to \ell^\infty(B), \qquad\qquad \tilde{\psi}(\lambda)(b) = \psi(b, \lambda(b))$$

   is a continuous map by lemma E.3. Theorem 4.2 implies $\sup_{b\in B}\|\hat{\lambda}_n(b) - \lambda(b)\| \xrightarrow{p} 0$, so the continuous mapping theorem implies $\tilde{\psi}(\hat{\lambda}_n) \xrightarrow{p} \tilde{\psi}(\lambda) = 0$ in $\ell^\infty(B)$; therefore

$$\sup_b \|\nabla^2_\lambda \nu(b, \hat{\lambda}_n(b)) - \nabla^2_\lambda \nu(b, \lambda(b))\| = \|\tilde{\psi}(\hat{\lambda}_n)\|_B \xrightarrow{p} \|\tilde{\psi}(\lambda)\|_B = 0$$

2. With probability approaching 1, the eigenvalues of $\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b))$ are strictly negative for every $b \in B$.

   As in the proof of lemma E.4, Weyl's perturbation theorem (Bhatia (1997) corollary III.2.6) implies uniform consistency of the eigenvalues of $\nabla_\lambda \hat{\nu}_n$:

$$\begin{aligned}&\sup_{b\in B} \max_k \left|\alpha_k\left(\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b))\right) - \alpha_k\left(\nabla^2_\lambda \nu(b, \lambda(b))\right)\right| \\ &\qquad\qquad \leq \sup_{b\in B} \left\|\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b)) - \nabla^2_\lambda \nu(b, \lambda(b))\right\| \xrightarrow{p} 0\end{aligned}$$

   As in lemma F.6, continuity of $b \mapsto \alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right)$ and compactness of B implies through the extreme value theorem that $\sup_{b\in B} \alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right) = \max_{b\in B} \alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right) < 0$. Let $\varepsilon = -\max_{b\in B} \alpha_1\left(\nabla^2_\lambda \nu(b, \lambda(b))\right)/2$ and notice $\varepsilon > 0$. With probability approaching one $\sup_{b\in B}\left\|\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b)) - \nabla^2_\lambda \nu(b, \lambda(b))\right\| < \varepsilon$, and on this event $\sup_{b\in B} \alpha_1\left(\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b))\right) \leq$

$-\varepsilon < 0$. Since $\hat{\nu}_n(b,\lambda)$ is concave in λ, the eigenvalues of $\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b))$ are all strictly negative for all $b \in B$.

When $\sup_{b\in B} \alpha_1 \left(\nabla^2_\lambda \hat{\nu}_n(b, \hat{\lambda}_n(b))\right) < 0$, the implicit function theorem implies there exists a neighborhood of b such that $\hat{\lambda}_n(b)$ is the unique solution to (52) and continuous on that neighborhood. As this holds for all $b \in B$, $\hat{\lambda}_n : B \to \mathbb{R}^{d_g+K}$ is a well defined, continuous function.

This occurs with probability approaching one, so $P(\sqrt{n}(\hat{\nu}_n - \nu) \in \mathcal{C}(B)) \to 1$ and as argued above the Portmanteau theorem implies $P(\mathbb{G}_\nu \in \mathcal{C}(B)) = 1$. □

Theorem 4.7 (Asymptotic distribution of the breakdown point)**.** *Suppose assumptions 1, 2, 3, and 4 hold. Then*

$$\sqrt{n}(\hat{\delta}^{BP}_n - \delta^{BP}) \xrightarrow{L} \inf_{b\in \boldsymbol{m}(\nu)} \mathbb{G}_\nu(b)$$

where $\boldsymbol{m}(\nu) = \arg\min_{b\in B\cap \boldsymbol{B}_0} \nu(b)$.

Proof. Lemma E.5 shows that

$$\iota : \ell^\infty(B) \to \mathbb{R}, \qquad \iota(\nu) = \inf_{b\in B\cap \mathbf{B}_0} \nu(b)$$

is Hadamard directionally differentiable tangentially to $\mathcal{C}(B) = \{f : B \to \mathbb{R}\ ;\ f \text{ is continuous}\}$ at any $\nu \in \mathcal{C}(B)$, with

$$\iota'_\nu(h) = \inf_{b\in \mathbf{m}(\nu)} h(b), \qquad \text{where} \qquad \mathbf{m}(\nu) = \arg\min_{b\in B\cap \mathbf{B}_0} \nu(b)$$

So it suffices to verify the following two assumptions of Fang & Santos (2019):

1. (i) requires ι to be a function mapping a subset of a Banach space to another Banach space. $(\ell^\infty(B), \|\cdot\|_B)$ and $(\mathbb{R}, |\cdot|)$ are indeed Banach spaces.

   (ii) requires that ι is Hadamard directionally differentiable at ν tangentially to $\mathcal{C}(B)$. This was shown in lemma E.5.

2. (i) requires $\nu \in \mathcal{C}(B)$, $\hat{\nu}_n$ is a mapping of data to $\ell^\infty(B)$ such that $\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu$ in $\ell^\infty(B)$. Lemma F.4 shows $\nu \in \mathcal{C}(B)$, theorem 4.6 implies $\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu$.

   (ii) requires that $\mathbb{G}_\nu$ be tight and supported on a subset of $\mathcal{C}(B)$. Tightness of $\mathbb{G}_\nu$ is implied by theorem 4.6, and lemma F.9 shows that $P(\mathbb{G}_\nu \in \mathcal{C}^\infty(B)) = 1$.

It follows by Fang & Santos (2019) theorem 2.1 that $\sqrt{n}(\iota(\hat{\nu}_n) - \iota(\nu)) = \iota'_\nu(\sqrt{n}(\hat{\nu}_n - \nu)) + o_p(1)$, and so $\sqrt{n}(\hat{\delta}^{BP}_n - \delta^{BP}) \xrightarrow{L} \iota'_\nu(\mathbb{G}_\nu)$ in $\mathbb{R}$. □

F.2.2 Bootstrap

Theorem 4.8 (Score bootstrap consistency)**.** *Suppose assumptions 1, 2, 3, and 4 hold, and* $\{W_i\}_{i=1}^n$ *satisfies assumption 5. Let* $\hat{\Phi}_n(b) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b))$*, and* $\hat{G}^*_n : B \to \mathbb{R}^{d_g+K+2}$ *be defined pointwise as*

$$\hat{G}^*_n(b) = \hat{\Phi}_n(b)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b)) \qquad (14)$$

Further define $\hat{\theta}_n^*(b) = \frac{1}{\sqrt{n}}\hat{G}_n^*(b) + \hat{\theta}_n(b)$. *Then conditional on* $\{D_i, D_iY_i, X_i\}_{i=1}^n$,

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \xrightarrow{L} \mathbb{G}$$

in outer probability.

Proof. Let $\mathrm{BL}_1\left(\ell^\infty(B \times \mathcal{I})\right)$ be the set of real functions f on $\ell^\infty(B \times \mathcal{I})$ whose absolute value and lipschitz constant are bounded by 1: $\sup_{h \in \ell^\infty(B\times\mathcal{I})}|f(h)| \leq 1$ and $|f(h_1) - f(h_2)| \leq \|h_1 - h_2\|_{B\times\mathcal{I}}$ for all $h_1, h_2 \in \ell^\infty(B \times \mathcal{I})$. By van der Vaart & Wellner (1997) theorem 1.12.4, it suffices to show

$$\sup_{f \in \mathrm{BL}_1(\ell^\infty(B\times\mathcal{I}))} \left| E\left[f\left(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)\right) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right] - E\left[f(\mathbb{G})\right] \right| \xrightarrow{p^*} 0$$

Define

$$G_n^* = \Phi(b)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i \phi(D_i, D_iY_i, X_i, b, \theta_0(b))$$

and notice that

$$\sup_{f \in \mathrm{BL}_1(\ell^\infty(B\times\mathcal{I}))} \left| E\left[f(\hat{G}_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right] - E\left[f(\mathbb{G})\right] \right|$$

$$\leq \sup_{f \in \mathrm{BL}_1(\ell^\infty(B\times\mathcal{I}))} \left| E\left[f(\hat{G}_n^*) - f(G_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right] \right| \tag{53}$$

$$+ \sup_{f \in \mathrm{BL}_1(\ell^\infty(B\times\mathcal{I}))} \left| E\left[f(G_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right] - E\left[f(\mathbb{G})\right] \right| \tag{54}$$

The second term (54) can be controlled through van der Vaart & Wellner (1997) theorem 2.9.6. First note that van der Vaart & Wellner (1997) problem 2.9.1 shows that if $E[|W_i|^{2+a}] < \infty$ for some $a > 0$, then $\|W_i\|_{2,1} = \int_0^\infty \sqrt{P(|W_i| > x)}dx < \infty$ (see also p. 177). The proof of theorem 4.6 shows that $\left\{\Phi(b)^{-1}\phi(d, dy, x, b, \theta_0(b)) \; ; \; b \in B\right\}$ is Donsker. van der Vaart & Wellner (1997) theorem 2.9.6 then implies

$$\sup_{f \in \mathrm{BL}_1(\ell^\infty(B\times\mathcal{I}))} \left| E\left[f(G_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right] - E\left[f(\mathbb{G})\right] \right| \xrightarrow{p^*} 0$$

The first term (53) requires more argument. For legibility, the proof that this term is $o_p(1)$ is broken into steps.

1. Establish that $\sup_{b \in B}\|\hat{G}_n^*(b) - G_n^*(b)\| = o_p(1)$.

   First notice that, just as in lemma F.8, the mean value theorem (Coleman (2012) Corollary 3.2) implies

$$\|w\phi(d, dy, x, b_1, \theta_1) - w\phi(d, dy, x, b_2, \theta_2)\|$$

$$\leq |w| \left[\sup_{(b,\theta) \in \Theta^B} \left\| \nabla_{(b,\theta)} \phi(d, dy, x, b, \theta) \right\|_o \right] \|(b_1, \theta_1) - (b_2, \theta_2)\|$$

and so, where F is the envelope from assumption 4 (v),

$$\begin{aligned}
&|W|\left[\sup_{(b,\theta)\in\Theta^B}\left\|\nabla_{(b,\theta)}\phi(D,DY,X,b,\theta)\right\|_o\right]\leq |W|F(D,DY,X),\\
\Longrightarrow\ & E\left[W^2F(D,DY,X)^2\right]=E[W^2]E\left[F(D,DY,X)^2\right]=E\left[F(D,DY,X)^2\right]<\infty
\end{aligned}$$

thus $\left\{w\phi(d,dy,x,b,\theta)\ ;\ (b,\theta)\in\Theta^B\right\}$ is a special case of van der Vaart (2007) example 19.7, and hence Donsker. The CMT then implies $\sup_{(b,\theta)\in\Theta^B}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(D_i,D_iY_i,X_i,b,\theta)\right\|=O_p(1)$. Lemma F.7 shows $\sup_{b\in B}\left\|\hat{\Phi}_n(b)^{-1}-\Phi(b)^{-1}\right\|=o_p(1)$, hence

$$\begin{aligned}
&\sup_{b\in B}\left\|\left[\hat{\Phi}_n(b)^{-1}-\Phi_n(b)^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(D_i,D_iY_i,X_i,b,\hat{\theta}_n(b))\right\|\\
&\qquad\leq\sup_{b\in B}\left\|\hat{\Phi}_n(b)^{-1}-\Phi_n(b)^{-1}\right\|\sup_{(b,\theta)\in\Theta^B}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(D_i,D_iY_i,X_i,b,\theta)\right\|\\
&\qquad\xrightarrow{p}0
\end{aligned}\tag{55}$$

Finally, the conditions of lemma E.6 are satisfied: $\hat{\theta}_n,\theta_0:B\to\mathbb{R}^{d_g+K}$ with $\hat{\theta}_n(b),\theta_0(b)\in\Theta^b$ for each b,

(i) $\Theta^B=\{(b,\theta)\ ;\ b\in B,\ \theta\in\Theta^b\}$ is compact,

(ii) $(b,\theta)\mapsto w\phi(d,dy,x,b,\theta)$ is continuous,

(iii) $\sup_{b\in B}\left\|\hat{\theta}_n(b)-\theta_0(b)\right\|=o_p(1)$, and

(iv) $\left\{w\phi(d,dy,x,b,\theta)\ ;\ (b,\theta)\in\Theta^B\right\}$ is Donsker with square integrable envelope $|w|F(d,dy,x)$, and $E[W\phi(D,DY,X,b,\theta)]=E[W]E[\phi(D,DY,X,b,\theta)]=0$,

which implies

$$\sup_{b\in B}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(D_i,D_iY_i,X_i,b,\hat{\theta}_n(b))-\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(D_i,D_iY_i,X_i,b,\theta_0(b))\right\|=o_p(1)\tag{56}$$

by lemma E.6. Now use (55), (56), and $\sup_{b\in B}\|\Phi(b)^{-1}\| < \infty$ as shown in lemma F.6 to see

$$\begin{aligned}
&\sup_{b\in B}\|\hat{G}_n^*(b) - G_n^*(b)\| \\
&\quad = \sup_{b\in B}\left\|\hat{\Phi}_n(b)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b)) - \Phi(b)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \theta_0(b))\right\| \\
&\quad \leq \sup_{b\in B}\left\|\left[\hat{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b))\right\| \\
&\qquad + \sup_{b\in B}\left\|\Phi(b)^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \theta_0(b))\right]\right\| \\
&\quad \leq \sup_{b\in B}\left\|\hat{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right\|\sup_{b\in B}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b))\right\| \\
&\qquad + \sup_{b\in B}\|\Phi(b)^{-1}\|\sup_{b\in B}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(D_i, D_iY_i, X_i, b, \theta_0(b))\right\| \\
&\quad = o_p(1) + o_p(1)
\end{aligned}$$

2. The term in (53) converges in probability to zero.

   First note that for any $f \in \mathrm{BL}_1(\ell^\infty(B\times\mathcal{I}))$, $|f(h_1) - f(h_2)| \leq \min\{2, \|h_1 - h_2\|_{B\times\mathcal{I}}\}$. So for any $\eta \in (0, 2]$,

$$\begin{aligned}
&\left|E\left[f(\hat{G}_n^*) - f(G_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right]\right| \leq E\left[\left|f(\hat{G}_n^*) - f(G_n^*)\right| \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right] \\
&\leq P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right) \\
&\qquad\qquad \times E\left[\min\{2, \|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}}\} \mid \{D_i, D_iY_i, X_i\}_{i=1}^n, \|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta\right] \\
&\quad + P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} \leq \eta \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right)\eta \\
&\leq 2P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right) + \eta
\end{aligned}$$

   The upper bound doesn't depend on f, hence

$$\begin{aligned}
&\sup_{f\in BL_1(\ell^\infty(B\times\mathcal{I}))}\left|E\left[f(\hat{G}_n^*) - f(G_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right]\right| \\
&\qquad\qquad \leq 2P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right) + \eta \qquad (57)
\end{aligned}$$

   Markov's inequality implies that for any $\varepsilon > 0$,

$$\begin{aligned}
&P\left(P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right) > \varepsilon\right) \\
&\qquad \leq \frac{1}{\varepsilon}E\left[P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right)\right] \\
&\qquad \leq \frac{1}{\varepsilon}P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta\right) \qquad (58)
\end{aligned}$$

   Step 1 established $\sup_{b\in B}\|\hat{G}_n(b) - G_n^*(b)\| = o_p(1)$, so $P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta\right) \to 0$ and hence (58) implies $P\left(\|\hat{G}_n^* - G_n^*\|_{B\times\mathcal{I}} > \eta \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right) = o_p(1)$. Since $\eta \in (0,2]$ was arbitrary, (57) implies $\sup_{f\in BL_1(\ell^\infty(B\times\mathcal{I}))}\left|E\left[f(\hat{G}_n^*) - f(G_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right]\right| = o_p(1)$.

Thus, the term in (53) is $o_{p^*}(1)$ and the term in (54) is $o_p(1)$, giving the result that

$$\sup_{f\in \text{BL}_1(\ell^\infty(B\times\mathcal{I}))}\left|E\left[f(\hat{G}_n^*) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right] - E\left[f(\mathbb{G})\right]\right| \xrightarrow{p^*} 0$$

□

Corollary F.10 (Bootstrap consistency for the value function)**.** *Suppose assumptions 1, 2, 3, and 4 hold, and $\{W_i\}_{i=1}^n$ satisfies assumption 5. Let $\hat{G}_n^*$ be defined as in theorem 4.8, and $\hat{\nu}_n^* : B \to \mathbb{R}$ defined pointwise by $\hat{\nu}_n^*(b) = \frac{1}{\sqrt{n}}\hat{G}_n^*(b,1) + \hat{\nu}_n(b)$. Then conditional on $\{D_i, D_iY_i, X_i\}_{i=1}^n$,*

$$\sqrt{n}(\hat{\nu}_n^* - \hat{\nu}_n) \xrightarrow{L} \mathbb{G}_\nu$$

in outer probability.

Proof. As in theorem 4.8, it suffices to show

$$\sup_{f\in \text{BL}_1(\ell^\infty(B))}\left|E\left[f\left(\sqrt{n}(\hat{\nu}_n^* - \hat{\nu}_n)\right) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right] - E\left[f(\mathbb{G}_\nu)\right]\right| \xrightarrow{p^*} 0$$

For any $h \in \ell^\infty(B\times\mathcal{I})$, $h_1(b) = h(b,1)$ defines an element h_1 of $\ell^\infty(B)$. For any $f \in \text{BL}_1(\ell^\infty(B))$, there exists $\tilde{f} \in \text{BL}_1(\ell^\infty(B\times\mathcal{I}))$ such that $\tilde{f}(h) = f(h_1)$. Let $\widetilde{\text{BL}}_1(\ell^\infty(B\times\mathcal{I})) \subset \text{BL}_1(\ell^\infty(B\times\mathcal{I}))$ collect such $\tilde{f}$ functions, and notice that

$$\begin{aligned}
&\sup_{f\in \text{BL}_1(\ell^\infty(B))}\left|E\left[f\left(\sqrt{n}(\hat{\nu}_n^* - \hat{\nu}_n)\right) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right] - E\left[f(\mathbb{G}_\nu)\right]\right| \\
&\qquad = \sup_{\tilde{f}\in \widetilde{\text{BL}}_1(\ell^\infty(B\times\mathcal{I}))}\left|E\left[\tilde{f}\left(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)\right) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right] - E\left[\tilde{f}(\mathbb{G})\right]\right| \\
&\qquad \leq \sup_{f\in \text{BL}_1(\ell^\infty(B\times\mathcal{I}))}\left|E\left[f\left(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)\right) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right] - E\left[f(\mathbb{G})\right]\right| \\
&\qquad \xrightarrow{p^*} 0
\end{aligned}$$

□

F.2.3 Confidence intervals

Lemma F.11 (Plug in bootstrap confidence interval consistency)**.** *Suppose assumptions 1, 2, 3, and 4 hold, $\{W_i\}_{i=1}^n$ satisfies assumption 5, and $\arg\min_{b\in B\cap \boldsymbol{B}_0}\nu(b)$ is unique. Let*

$$\hat{c}_{1-\alpha,n}^{Plug} = \inf\left\{c\,;\; P\left(\sqrt{n}(\iota(\hat{\nu}_n^*) - \iota(\hat{\nu}_n)) \leq c \mid \{D_i, D_iY_i, X_i\}_{i=1}^n\right) \geq \alpha\right\}$$

then

$$\lim_{n\to\infty} P\left(\hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}}\hat{c}_{1-\alpha,n}^{Plug} \leq \delta^{BP}\right) = 1-\alpha$$

Proof. First apply Fang & Santos (2019) theorem 3.1. Fang & Santos (2019) assumptions 1 and 2 are verified in the proof of theorem 4.7. Assumption 3 is about the bootstrap procedure for the value function:

3. (i) $\hat{\nu}_n^* : \{D_i, D_iY_i, X_i, W_i\}_{i=1}^n \to \ell^\infty(B)$ with $\{W_i\}_{i=1}^n$ independent of $\{D_i, D_iY_i, X_i\}_{i=1}^n$.

   (ii) $\hat{\nu}_n^*$ satisfies

$$\sup_{f \in \mathrm{BL}_1(\ell^\infty(B))} \left| E\left[f\left(\sqrt{n}(\hat{\nu}_n^* - \hat{\nu}_n) \right) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right] - E\left[f(\mathbb{G}_\nu) \right] \right| \xrightarrow{p} 0$$

   (iii) $\sqrt{n}(\hat{\nu}_n^* - \hat{\nu}_n)$ is asymptotically measurable jointly in $\{D_i, D_iY_i, X_i, W_i\}_{i=1}^n$

   (iv) $f(\sqrt{n}(\hat{\nu}_n^* - \hat{\nu}_n))$ is a measurable function of $\{W_i\}_{i=1}^n$ outer almost surely in $\{D_i, D_iY_i, X_i\}_{i=1}^n$ for any continuous and bounded $f : \ell^\infty(B) \to \mathbb{R}$.

Conditions (i) and (ii) are shown in theorem 4.6 and corollary F.10. Theorem 4.6 also shows $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) = \hat{G}_n^* = G_n^* + o_p(1)$. Since G_n^* converges weakly to a tight limit (unconditionally), so does $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$, and van der Vaart & Wellner (1997) lemma 1.3.8 implies $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ is asymptotically measurable. Lastly, note that because $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ is continuous in $\{W_i\}_{i=1}^n$, so is $f(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n))$, and hence $f(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n))$ is measurable in $\{W_i\}_{i=1}^n$. Finally, note that $\mathbb{G}_\nu$ is Gaussian and its support is $\mathcal{C}(B)$, a vector subspace of $\ell^\infty(B)$. Fang & Santos (2019) theorem 3.1 then implies

$$\sup_{f \in \mathrm{BL}_1(\mathbb{R})} \left| E\left[f\left(\sqrt{n}(\iota(\hat{\nu}_n^*) - \iota(\hat{\nu}_n)) \right) \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right] - E\left[f\left(\iota_\nu'(\mathbb{G}_\nu) \right) \right] \right| = o_p(1) \tag{59}$$

Since ι_ν' is linear, ι is fully Hadamard differentiable at ν and $\iota_\nu'(\mathbb{G}_\nu)$ is a Gaussian distribution on $\mathbb{R}$. The cumulative distribution function of $\iota_\nu'(\mathbb{G}_\nu)$ is continuous and strictly increasing at all points, so lemma E.7 implies that for any α,

$$\begin{aligned} P\left(\hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}} \hat{c}_{1-\alpha,n}^{Plug} \leq \delta^{BP} \right) &= P\left(\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) \leq \hat{c}_{1-\alpha,n}^{Plug} \right) \\ &= P\left(\sqrt{n}(\iota(\hat{\nu}_n) - \iota(\nu)) \leq \hat{c}_{1-\alpha,n}^{Plug} \right) \\ &\to 1 - \alpha \end{aligned}$$

□

Theorem 4.9 (Confidence interval consistency)**.** *Suppose assumptions 1, 2, 3, and 4 hold, and* $\{W_i\}_{i=1}^n$ *satisfies assumption 5. Let* $\hat{G}_n^*$ *be as defined in theorem 4.8, and* $\hat{\nu}_n^* : B \to \mathbb{R}$ *defined pointwise by* $\hat{\nu}_n(b) = \frac{1}{\sqrt{n}} \hat{G}_n(b, 1) + \hat{\theta}_n(b, 1)$*. Let* $\hat{b}_n^i$ *solve* $\min_{b \in B \cap \boldsymbol{B}_0} \hat{\nu}_n(b)$*, and*

$$\hat{c}_{1-\alpha,n}^{Simple} = \inf\left\{ c \,;\, P\left(\sqrt{n}(\hat{\nu}_n^*(\hat{b}_n^i) - \hat{\nu}_n(\hat{b}_n^i)) \leq c \mid \{D_i, D_iY_i, X_i\}_{i=1}^n \right) \geq 1 - \alpha \right\}$$

If $\boldsymbol{m}(\nu) = \arg\min_{b \in B \cap \boldsymbol{B}_0} \nu(b)$ *is the singleton* $\{b^i\}$*, then*

$$\lim_{n \to \infty} P\left(\hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}} \hat{c}_{1-\alpha,n}^{Simple} \leq \delta^{BP} \right) = 1 - \alpha$$

Proof. The proof uses the "consistent alternative" of Fang & Santos (2019), theorem 3.2. Since b^i uniquely solve $\min_{b\in B\cap\mathbf{B}_0}\nu(b)$, $\iota'_\nu(h)=h(b^i)$. Let $\hat{\iota}'_n(h)=h(\hat{b}^i_n)$ be the estimator of the derivative ι'_ν.

First notice that $\hat{b}^i_n$ is consistent for b^i. $B\cap\mathbf{B}_0$ is compact, ν is continuous by lemma F.4, and $\sup_{b\in B\cap\mathbf{B}_0}|\hat{\nu}_n(b)-\nu(b)|\leq\sup_{b\in B}|\hat{\nu}_n(b)-\nu(b)|\xrightarrow{p}0$ by theorem 4.2, so $\hat{b}^i_n\xrightarrow{p}b^i$ by standard extremum estimator arguments (e.g., Newey & McFadden (1994) theorem 2.1).

Next, observe that for any $h\in\mathcal{C}(B)$, $|\hat{\iota}'_n(h)-\iota'_\nu(h)|=|h(\hat{b}^i_n)-h(b^i)|=o_p(1)$ by the continuous mapping thoerem. Finally, for any $h_1,h_2\in\ell^\infty(B)$,

$$\left|\hat{\iota}'_n(h_1)-\hat{\iota}'_n(h_2)\right|=\left|h_1(\hat{b}^i_n)-h_2(\hat{b}^i_n)\right|\leq\sup_{b\in B}|h_1(b)-h_2(b)|=\|h_1-h_2\|_B$$

Thus Fang & Santos (2019) lemma S.3.6 implies Fang & Santos (2019) assumption 4 holds.

Assumption 1 and 2 of Fang & Santos (2019) is verified in the proof of theorem 4.7, and assumption 3 is verified in the proof of F.11. Thus Fang & Santos (2019) theorem 3.2 implies

$$\sup_{f\in\mathrm{BL}_1(\ell^\infty(B))}\left|E\left[f\left(\hat{\iota}'_n\left(\sqrt{n}(\hat{\nu}^*_n-\hat{\nu}_n)\right)\right)\mid\{D_i,D_iY_i,X_i\}_{i=1}^n\right]-E\left[f(\iota'_\nu(\mathbb{G}_\nu))\right]\right|=o_p(1)$$

Since $\iota'_\nu(\mathbb{G}_\nu)$ is Gaussian and hence continuous and strictly increasing at all points, lemma E.7 implies that for any α,

$$P\left(\hat{\delta}^{BP}_n-\frac{1}{\sqrt{n}}\hat{c}^{Simple}_{1-\alpha,n}\leq\delta^{BP}\right)=P\left(\sqrt{n}(\hat{\delta}^{BP}_n-\delta^{BP})\leq\hat{c}^{Simple}_{1-\alpha,n}\right)\to1-\alpha.$$

□