Testing for Moral Hazard When Adverse Selection is Present

Juan Carlos Escanciano* Indiana University

Bernard Salanié† Columbia University

Neşe Yıldız‡ University of Rochester

First version: May, 2009. This version: August 13, 2016.

Abstract

Unobservable heterogeneity in agents makes it notoriously difficult to test for moral hazard: since different agents choose different contracts, it is hard to isolate the causal impact of the contract on outcomes. We show how exogeneous variation in contract menus allows to test for moral hazard under selection of unknown form. Our test relies on independence between outcomes and the instruments that drive the exogeneous variation. We propose parametric, semiparametric and nonparametric testing procedures. We detail their construction and their properties in a model with two outcomes and two levels of effort. The proposed test statistics have smaller bias and are less sensitive to smoothing parameters than existing conditional mean independence tests. We also suggest another strategy that can be applied when the distribution of outcomes increases with effort.

JEL Codes: C14, C52, D82.
Keywords: Moral hazard; Testing contract theory; Insurance; Conditional mean independence.

*Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. Phone: +1 (812) 855 7925. Fax: +1 (812) 855 3736. E-mail: jescanci@indiana.edu.

†Department of Economics, Columbia University, 1022 International Affairs Building 420 W 118th Street New York, NY 10027, USA. E-mail: bsalanie@columbia.edu.

‡Correspondence: Department of Economics, University of Rochester, 231 Harkness Hall, Rochester, NY 14627; Email: nese.yildiz@rochester.edu; Phone: 585-275-5782; Fax: 585-256 23 09.

1 Introduction

Testing for the presence of moral hazard is a hard problem when agents are unobservably heterogeneous—what the contract literature calls "hidden information". This has been known for some time in the insurance context for instance, where several approaches have been used to circumvent this difficulty[1]. We propose in this paper a test for moral hazard (or lack thereof) that relies on the availability of an instrument that shifts contract choice but is excluded from the outcome equation. We allow for the presence of hidden information in the sense that the Agent can have a privately known type, that neither the Principal nor the econometrician can observe. To describe our contribution, we use the language of insurance models in this introduction. This is purely for expository purposes; nothing depends on it. The reader can simply substitute say "employment contract" for "insurance contract" to apply our results to labor economics for instance.

The basic model of insurance has a risk averse agent deciding how much coverage to buy against the risk of a loss. To fix ideas, we will illustrate this with car insurance; there the loss Y is caused by an accident, and different levels of coverage are available. The driver may be insured only if she is not at fault; or also if she is at fault, with various levels of deductibles and/or proportional reimbursements. We denote $R_d(Y)$ the transfer from the insurer to the insuree under contract d when the insuree incurs a loss Y, and P_d the premium the insuree pays. These are normally conditional on covariates (e.g. describing the driver as well as her car); we drop them from the notation here for simplicity. We denote the set of contracts that are available to this driver $\mathcal{D}$. Note for future reference that $\mathcal{D}$ can be represented as a set of pairs $(R_d(\cdot), P_d)$ indexed by d. It can vary with observed covariates, but not with the agent's private type $v \in \mathcal{V}$ since that is unobserved by the insurer.

Let an insuree with private type v have a Bernoulli utility index u_v, and start with an insurance relationship without moral hazard: this insuree faces an exogenous distribution of losses F_v. Note that v enters both the utility function (via initial wealth and attitude towards risk) and the distribution of losses. This insuree will choose her insurance contract d by solving

$$\max_{d \in \mathcal{D}} \int u_v(R_d(Y) - Y - P_d) dF_v(Y) \tag{1}$$

To illustrate: the standard Rothschild-Stiglitz model has two states. The loss can be 0, or some fixed $\bar{Y}$. There are two types of insurees, $v = L, H$, and the loss $\bar{Y}$ occurs with a probability p_v; so that F_v puts a mass p_v on $\bar{Y}$ and a mass $(1 - p_v)$ on 0. Finally, the two types have the same utility function $u_v \equiv u$.

We now introduce moral hazard in this model: the insuree may be able to influence the distribution of losses by exerting costly effort. A driver, for instance, may exercise more or less caution; and she may be influenced in her choice of effort by her insurance coverage. Let $e \in \mathcal{E}$ denote the effort, and $F_v(\cdot; e)$ the cdf of losses when the agent expends effort e. Then the agent chooses both

[1] See for instance the recent survey by Chiappori and Salanié (2014).

contract and effort by solving

$$\max_{d\in\mathcal{D},e\in\mathcal{E}} \int u_v(R_d(Y) - Y - P_d; e) dF_v(Y; e). \tag{2}$$

Note that utility depends both on final wealth and on effort now.

We can break down this problem further. For any choice of coverage d, the insuree chooses a level of effort $e = \bar{e}_v(d)$. This choice of effort depends both on the coverage d and, via utility and "technology", on the agent's type v. In a first stage, the insuree chooses her coverage by solving

$$\max_{d\in\mathcal{D}} \int u_v(R_d(Y) - Y - P_d; \bar{e}_v(d)) dF_v(Y; \bar{e}_v(d)), \tag{3}$$

with a solution $d = D_v^*(\mathcal{D})$. Plugging this into the choice of effort gives

$$e = \bar{e}_v\left(D_v^*(\mathcal{D})\right) \equiv e_v^*(\mathcal{D});$$

and realized losses will be drawn from the distribution

$$F_v\left(\cdot; e_v^*(\mathcal{D})\right) \equiv F_v^*(\cdot; \mathcal{D}).$$

This simple model shows why it is so difficult to test for moral hazard (or a fortiori to quantify it) in the presence of adverse selection. Suppose that the data available has both choice of coverage $d \in \mathcal{D}$ and realized losses Y. After conditioning on covariates, the econometrician will face data that shows observably identical insurees choosing different levels of coverage from a set $\mathcal{D}$ and having different losses Y. The corresponding data-generating process is

$$\begin{aligned} D_i &= D_{v_i}^*(\mathcal{D}) \\ Y_i &\sim F_{v_i}^*(\cdot; \mathcal{D}) \end{aligned}$$

where v_i is the unobserved type of insuree i. There is of course a connection between losses Y_i and coverage D_i: we know that $F_{v_i}^*(\cdot; \mathcal{D})$ is in fact $F_{v_i}\left(\cdot; \bar{e}_{v_i}(D_i)\right)$. But since the data is typically uninformative about effort choices, this is only very indirect information. It is easy to check that excluding some covariates from either u_v or F_v does not help either, as effort choice, contract choice and observed losses all depend on both utility and technology.

Now consider testing for moral hazard. In the absence of moral hazard, the distribution of losses would be an exogenous F_{v_i} as in (1); with moral hazard, it becomes a function of two additional variables, the choice of coverage D_i and the set of available contracts $\mathcal{D}$. If $\mathcal{D}$ does not vary in the data, testing for the null of no moral hazard involves testing that

$$Y_i \perp\!\!\!\perp D_i \mid v_i$$

while only observing the joint distribution of (Y_i, D_i). This would be easy if types v_i were observed; but it is clearly a hopeless task in the presence of adverse selection—or, more generally, if the econometrician only observes a subset of the payoff-relevant variables.

An alternative is to use variation in the menu of contracts $\mathcal{D}$. Recall that this is an indexed set of pairs, each of which has a premium and a reimbursement schedule. Any exogenous variation in the set of premia and/or reimbursement schedules can be used as an instrument: for any given agent v_i, it changes (or may change) the choice of coverage $D_i = D^*_{v_i}(\mathcal{D})$ but it does not affect the distribution of losses conditional on v_i and on the coverage chosen D_i. This is the approach to identification we use in this paper. We will show that if the change in the menu of contracts is truly exogenous, it allows the analyst to test for moral hazard in the presence of any kind of adverse selection, whether that bears on risk, on utility, or both[2].

In addition, we show that when contracts can be ordered so that (i) effort is monotonic with contract choice, conditional on type; (ii) contract choice is monotonic with respect to the instrument; and (iii) outcomes are monotonic in effort, then with moral hazard the distribution of outcomes must be stochastically monotonic with respect to the instrument. This provides us with an alternative testing strategies. Note that while assumptions (i)-(iii) are quite natural in the insurance context, it is easy to imagine settings in which they would be less credible.

Truly exogenous variation in contract choice is key to our results. Changes in the set of available contracts are often a response of insurers to perceived changes in demand. Such changes are typically not valid instruments. Randomized experimentation by insurers would, but it is a rare event[3]. On the other hand, several papers in insurance economics have exploited (for different purposes) variation in menus of contracts that is arguably exogenous. Cohen and Einav (2007) used both informal experimentation by insurers and inflation-driven adjustments to a nominal cap on deductibles. Einav et al (2013) rely on variation in the health insurance options offered to different groups of Alcoa workers at different points of time, stemming from staggered timing of new union contracts[4]. Handel (2013) observed workers' health insurance choices before and after a major change in the options offered by another large company. Regulatory changes can also generate plausible instruments, depending on their motivation and timing. Chiappori, Durand, and Geoffard (1998) exploited such an exogenous change in French health insurance: the replacement of full coverage with a 10 percent copayment in 1994. Dionne and Vanasse (1997) used changes in the definition of the "no fault" regime for car insurance in Québec. Annan (2015) uses a strategy related to ours to analyze the effect of a national reform of car insurance in Ghana. The reform made it harder to buy insurance on credit, making lower coverage more attractive. It led to a large reduction in claims, showing that moral hazard had played a large role in this market.

Given an exogenous instrument Z, a vector of covariates X, the contract choice D and the outcome Y, we provide conditions under which testing for the null of no moral hazard is tantamount to testing the conditional independence assumption

$$Y \perp\!\!\!\perp Z | X = x.$$

[2]A related but distinct idea exploits the dependence of contracts on observed characteristics that do not directly affect utility or technology. Weisburd (2013) applies this idea to data from a large Israeli firm which offers benefits based on occupation.

[3]See Manning et al (1987) for a study of the celebrated RAND experiment.

[4]As they explain clearly in section I.B of their paper, their "moral hazard" is really price elasticity—a common use of the term in health economics, but different from ours.

There exists an extensive literature on testing conditional independence; see for instance Su and White (2007, 2008) and references therein for a review of nonparametric testing results. We follow Delgado and González-Manteiga (2001, henceforth DG) in constructing nonparametric tests based on unconditional moments and restricted estimators. This choice has theoretical and practical advantages in terms of local power and only requires estimating low-dimensional nonparametric objects.

We give a complete discussion of our proposed test statistics and their properties in the common case in which both contract choice D and outcomes Y are binary; we normalize them to take values 0 and 1. Then we test for moral hazard as

$$Y \perp\!\!\!\perp P|X = x$$

with $P \equiv P(X, Z) = E(D|X, Z)$. In this *2-by-2 model*, we improve on the approach in DG by choosing test functions carefully. This allows us to obtain test statistics with smaller finite-sample bias. Our modification allows for optimal nonparametric first step estimators and data-driven optimal choices for bandwidths that are often used in practice (e.g. by cross-validation), and which are not permitted in DG's theory. Also, our nonparametric test statistic is less sensitive to smoothing parameters than alternative procedures.

We also complement our nonparametric tests with parametric and semiparametric tests that can be implemented by simple least squares methods in off-the-shelf statistical software, extending previous tests by Wooldridge (1990) to our setting.

Section 2 describes the basic testing idea. Section 3 investigates the 2-by-2 model in detail. It provides conditions under which our test has power; it proposes parametric and nonparametric test statistics under different specifications and establishes their asymptotic properties. Section 4 proposes a different testing strategy that relies on monotonicity in instruments and stochastic dominance restrictions. An Appendix gathers computational aspects of our nonparametric tests, and mathematical proofs for the main inference results.

2 The basic testing idea

Let us translate the economic model presented in the introduction into a statistical model. Consider the following two equations:

$$Y = g(e(D, X, V), X, V, \eta), \tag{4}$$
$$D = h(Z, X, V), \tag{5}$$

where

- Y denotes the outcome(s) of interest.
- X contains observed factors that influence the agent's choice of effort; they will normally also affect both the probability of the bad state and the choice of coverage.

- Z are additional observed variables that do not affect the choice of effort for a given coverage, but may affect contract choice by varying the set of available contracts.
- V is an unobservable representing the individual's type.
- η represents factors that influence the probability of a bad state conditional on effort.

The term V is the agent's private information. Note that while V is observed by the agent and determines her choice of coverage, η typically contains both information that accrues to the agent between the time she chooses coverage and the time she chooses effort, and shocks that are unobservable by the agent.
We make the following assumption on the variable Z:

Assumption 2.1 (Instrument Validity at x) *The pair (V, η) is independent of Z conditional on $X = x$.*

In order to proceed, we need to specify a null hypothesis. This is tantamount to defining "moral hazard", or its absence. According to the textbook definition (e.g. Salanié 2005, p. 119), moral hazard arises when the Agent can make unobservable decisions that affect the joint surplus of her interaction with the Principal, and their incentives are not aligned. In a model of insurance, the unobservable decision is effort. The marginal benefit of effort is a reduction in losses, which benefits both parties in a way that depends on the insurance contract; and the marginal cost of effort is born by the insuree only. This suggests the following null hypothesis:

For fixed x, effort does not change the distribution of losses;
that is, the function $(d, v) \to e(d, x, v)$ is constant.

On the other hand, this is stronger than it need be. Suppose that effort does not respond to contract choice; then the insurer does not have to worry that providing more coverage will increase claims, and moral hazard is irrelevant. This corresponds to a weaker null hypothesis:

For fixed x, effort does not depend on the insurance contract;
that is, the function $(d, v) \to e(d, x, v)$ only depends on v.

We use this weaker definition in what follows, and we specify:

$$(H_0) \text{ at } x : \text{the function } (d, v) \to e(d, x, v) \text{ only depends on } v.$$

Under this null hypothesis, outcomes are generated by

$$Y = g(e(X, V), X, V, \eta),$$

and given Assumption 2.1, Y is independent of Z conditional on $X = x$. We have established that

Proposition 2.1 *Under the null, if Assumption 2.1 holds at x then*

$$Y \perp\!\!\!\perp Z | X = x.$$

At this stage nothing in our assumptions guarantees that the reciprocal to Proposition 2.1 holds, i.e. that lack of conditional independence implies moral hazard. If for instance the contract choice does not vary at all with the instrument Z, the test will be useless since conditional independence will hold even with moral hazard. We therefore need to assume that instruments can pick up deviations from (H_0). To put it loosely, we want to find two possible values of Z that lead to different choices of effort, which in turn translate into different distributions of outcomes. The following assumption spells this out rigorously:

Assumption 2.2 (Instrument relevance at x) *There exist two disjoint subsets $\mathcal{Z}_1$ and $\mathcal{Z}_2$ of* $\operatorname{supp} Z|X = x$ *such that*

(i) $\Pr(Z \in \mathcal{Z}_j | X = x) > 0$ *for* $j = 1, 2$

(ii) for any v in $\operatorname{supp} V|X = x$*, denote $\mathcal{E}_j(v)$ the set $e(h(\mathcal{Z}_j, x, v), x, v)$ for $j = 1, 2$. Then for all measurable selections $e_1(v) \in \mathcal{E}_1(v)$ and $e_2(v) \in \mathcal{E}_2(v)$,*

$$\Pr(g(e_1(V), x, V, \eta) \neq g(e_2(V), x, V, \eta) \,|X = x) > 0.$$

The implications of Assumption 2.2 are obvious:

Proposition 2.2 (Power at x) *Let Assumption 2.2 hold at x. Then if (H_0) does not hold at x, Y and Z are not independent conditional on $X = x$.*

This proposition highlights that power comes from a combination of the richness of the variation in contract choice induced by instruments (via h) and the strength of the assumptions that can be imposed on the choice of effort (via the function e) and on the technology (via g). Note that in the common case when instruments only take two values z_1 and z_2, a test of Assumption 2.2 is feasible if we observe some agents repeatedly *under both instrument values*, and their unobserved heterogeneity (v, η) does not change[5]. If such an agent always has the same outcomes under z_1 and under z_2, this would suggest that Assumption 2.2 does not apply to him/her.

In the next section, we study a more specialized model; we will propose test statistics and also discuss the power of the corresponding test procedures.

[5]Or at least it can be matched across the two instrument values.

3 The 2-by-2 model: binary outcome and binary contract choice

To illustrate our approach, we will now focus on the simplest possible model in which moral hazard can arise: the menu of contracts $\mathcal{D}$ only has two elements, $D = 0, 1$; and outcomes Y only take two values which we also denote 0 and 1. In addition, we assume single-index monotonicity:

Assumption 3.1 (Monotonicity) *There is a function P such that $D = 1(V \leq P(X, Z))$.*

Then the *2-by-2 model* is

$$Y = 1(\eta \leq e(D, X, V)), \tag{6}$$
$$D = 1(V \leq P(X, Z)), \tag{7}$$

where henceforth $1(A)$ denotes the indicator function of the event A, i.e. $1(A) = 1$ if A is true, and zero otherwise. Without loss of generality, we can normalize V and η to be distributed as $U[0, 1]$ conditional on $X = x$. In particular, $P(X, Z) \equiv E(D|X, Z)$.

In this 2-by-2 model, under Assumption 2.1

$$\Pr(Y = 1|X = x, Z = z) = \int_0^{P(x,z)} F_{\eta|V,X}\left(e(1, x, v)|v, x\right) dv + \int_{P(x,z)}^1 F_{\eta|V,X}\left(e(0, x, v)|v, x\right) dv,$$

which only depends on z through $P(x, z)$. Therefore we only need to test that

$$Y \perp\!\!\!\perp P(X, Z)|X = x$$

and we can expect that such tests will have more power to detect moral hazard if the model is well-specified.[6] To simplify notation, we denote by P the random variable $P(X, Z)$.

3.1 Power in the 2-by-2 model

We provide sufficient conditions for our testing procedure to have power at x in this model, where the outcome is given by

$$Y = 1\left(\eta \leq e(1(V \leq P), x, V)\right).$$

Assumption 3.2 (Power at x in the 2-by-2 model)
There exist two disjoint subsets $\mathcal{Z}_1$ and $\mathcal{Z}_2$ of $\operatorname{supp} Z|X = x$ *such that*

(i) $\Pr(Z \in \mathcal{Z}_j|X = x) > 0$ *for* $j = 1, 2$.

[6] Conversely, if the test gives very different results when we use Z or $P(x, Z)$ this would point towards misspecification.

(ii) $\overline{p}_1(x) < \underline{p}_2(x)$, *where for* $j = 1, 2$, $\overline{p}_j(x) = \sup\{P(x,z) : z \in \mathcal{Z}_j\}$, *and* $\underline{p}_j(x) = \inf\{P(x,z) : z \in \mathcal{Z}_j\}$.

(iii) *For* $v \in \mathcal{V}_M(x) \equiv [\underline{p}_1(x), \overline{p}_2(x)]$, $sign(e(1,x,v) - e(0,x,v))$ *does not depend on* v, *where* $sign(t) = 1(t \geq 0) - 1(t < 0)$.

(iv) *For all* $v \in \mathcal{V}_m(X) \equiv (\overline{p}_1(X), \underline{p}_2(X))$, η *has full support on* $[0,1]$ *conditional on* $(X = x, V = v)$.

(v) $\Pr(e(1,x,V) \neq e(0,x,V) | X = x, V \in \mathcal{V}_m(X)) > 0$.

Proposition 3.1 *Suppose that* (H_0) *does not hold at* x. *Then under Assumptions 2.1, 3.1 and 3.2,* Y *and* Z *are not independent conditional on* $X = x$.

Proof: Let $z_1 \in \mathcal{Z}_1, z_2 \in \mathcal{Z}_2$. First note that for any value of v not in $\mathcal{V}_M(x)$, the choices of effort coincide under z_1 and z_2: e.g. for $v < \underline{p}_1(x)$ we have $e(1(V \leq P(x,z_1)), x, v) = e(1, x, v) = e(1(V \leq P(x,z_2)), x, v)$. Now using (iii) of Assumption 3.2, suppose that $e(1,x,v) \leq e(0,x,v)$ for each $v \in \mathcal{V}_M(x)$. For any such value of v,

$$e\left(1(v \leq P(x,z_1)), x, v\right) \geq e\left(1(v \leq \overline{p}_1(x)), x, v\right)$$

and

$$e\left(1(v \leq P(x,z_2)), x, v\right) \leq e\left(1(v \leq \underline{p}_2(x)), x, v\right).$$

Given Assumption 3.2.(ii), this implies that for any $v \in \mathcal{V}_M(x)$,

$$e\left(1(v \leq P(x,z_1)), x, v\right) > e\left(1(v \leq P(x,z_2)), x, v\right).$$

and therefore that

$$\begin{aligned}
&\Pr(Y = 1 | Z = z_1, X = x) - \Pr(Y = 1 | Z = z_2, X = x) \\
&= E\left[1(V \in \mathcal{V}_M(x)) \right. \\
&\times \left. \left(\Pr\left(\eta \leq e\left(1(V \leq P(x,z_1)), x, V\right) | V = v\right) - \Pr\left(\eta \leq e\left(1(V \leq P(x,z_2)), x, V\right) | V = v\right)\right) | X = x\right] \\
&< 0.
\end{aligned}$$

The argument is the same if $e(1,x,v) \leq (e(0,x,v)$ for each $v \in \mathcal{V}_M(x)$. Therefore Y and Z are not independent conditional on $X = x$. $\square$

If for instance the instrument Z is discrete, then we just choose two points z_1 and z_2 in its support at $X = x$ with $P(x,z_2) > P(x,z_1)$ and take $\mathcal{Z}_1 = \{z_1\}$ and $\mathcal{Z}_2 = \{z_2\}$. Our test will have power if $D = 1$ always induces more (or always induces less) effort than $D = 0$; or if V and η are independent and $D = 0$ and $D = 1$ induce different average efforts. If Z is continuous, then one can take small neighborhoods of such points z_1 and z_2 in the interior of the support of Z.

3.2 Testing in the 2-by-2 model

In the remainder of the paper we focus on testing the hypothesis[7]

$$Y \perp\!\!\!\perp P|X = x. \tag{8}$$

We also focus on the case where X is a continuous variable and we aim to test for (8) for almost surely (a.s) all x in the support of X, say $\mathcal{X}$. If X is in fact discrete and the researcher is interested in testing (8) at just one $x \in \mathcal{X}$, then the testing problem is simplified. We discuss the discrete covariate case in Remark 3.1 (p. 11).

The conditional independence $Y \perp\!\!\!\perp P|X$ can be characterized by the conditional moment restrictions

$$\Pr(Y = 1|P, X) = \Pr(Y = 1|X) \text{ a.s.}, \tag{9}$$

or equivalently since $Y \in \{0, 1\}$: $E(Y|P, X) = E(Y|X)$ a.s., or

$$E(\varepsilon|P, X) = 0 \text{ a.s.},$$

where $\varepsilon = Y - E(Y|X)$. These conditional moment restrictions in turn can be rewritten as the set of unconditional moment restrictions

$$R(\phi) := E(\varepsilon\phi(X, P)) = 0,$$

for all measurable functions ϕ such that the moment function $R(\phi)$ exists.

Let the data be the random sample $\{W_i = (Y_i, D_i, X_i, Z_i)\}_{i=1}^n$. The previous characterization suggests rejecting (9) for "large" values of the sample moments

$$R_n^0(\phi) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i\phi(X_i, P_i).$$

This is not a feasible statistic since neither P_i nor $E(Y|X)$ are observed; we construct a feasible version R_n by replacing these two quantities with estimators. Denote

- $\hat{Y}_{ni} = \hat{E}_n(Y_i|X_i)$ an estimator of $E(Y_i|X_i)$
- $\hat{P}_{ni}$ an estimator of $P_i = P(X_i, Z_i)$

where $\hat{Y}_{ni}$ and $\hat{P}_{ni}$ could be parametric, semiparametric or nonparametric fits.

We compute the residual $\hat{\varepsilon}_{ni} := Y_i - \hat{Y}_{ni}$ and we form the feasible sample moment functions

$$R_n(\phi) = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_{ni}\phi(X_i, \hat{P}_{ni}).$$

[7] Our results also apply to tests of $Y \perp\!\!\!\perp Z|X = x$, with the simplification that unlike P, Z is observed.

We call this test *parametric* when we use finitely many test functions ϕ. A parametric test rejects the null hypothesis when the quadratic form

$$S_n = nR_n'(\phi)\Omega_n^{-1}R_n(\phi), \tag{10}$$

takes large values (A' denotes the transpose of the matrix A). In this expression, ϕ denotes an m-dimensional vector of test functions and Ω_n estimates consistently the asymptotic variance

$$\Omega := E\left(\varepsilon_i^2\phi(X_i,P_i)\phi'(X_i,P_i)\right).$$

We will show in section 3.3 how to choose test functions so that the test statistic S_n converges under the null to a $\chi^2(m)$. We also will adapt the parametric approach of Wooldridge (1990) to show how this test can be implemented with routine least squares techniques.

Remark 3.1 *If X is discrete and one is interested in testing (8) at just one $x \in \mathcal{X}$, then the natural estimator for $E(Y_i|X_i = x)$ is $\hat{Y}_{ni} \equiv \hat{E}_n(Y_i|X_i = x) = n_x^{-1}\sum_{i=1}^n Y_i 1(X_i = x)$, where $n_x = \sum_{i=1}^n 1(X_i = x)$. The parametric test above applies with $\phi = (\phi_1, ..., \phi_m)'$, where the j-th component is $\phi_j(X_i, P_i) = \delta_j(P_i)1(X_i = x)$, for some measurable functions $\delta_j(\cdot)$, $j = 1, ..., m$.*

When an infinite family of test functions is used, we will call our test *nonparametric*. We will propose in Section 3.5 a nonparametric test where the test functions ϕ are indexed by $s \equiv (x, p) \in \mathcal{S}$, a compact subset of $\mathbb{R}^{d+1}$ where d is the dimension of X. We will define

$$R_n(s) = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_{ni}\phi_s(X_i, \hat{P}_{ni}) \tag{11}$$

the corresponding sample moments, now a stochastic process in $s \in \mathcal{S}$. The nonparametric test statistic is given by the quadratic form

$$C_n = n\int_{\mathcal{S}} \left(\frac{R_n(s)}{\sigma_n}\right)^2 d\mu(s), \tag{12}$$

where σ_n^2 is a consistent estimator of $E(\varepsilon^2)$, for example

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_{ni}^2,$$

and μ is a suitable integrating measure. Section 3.5 discusses a specific choice of (ϕ_s) and μ that leads to a closed form expression for C_n (see the Appendix for explicit computation of the nonparametric test statistic). Since the asymptotic null distribution of the nonparametric test is non-pivotal, we will also show how its critical values can be approximated by a simple multiplier bootstrap procedure.

Relative to other existing tests for conditional independence such as DG's, our choice of test functions has several advantages: it results in tests with a smaller finite-sample bias; it makes optimal nonparametric estimation of $E(Y_i|X_i)$ possible when X is low-dimensional; and it is less sensitive to the choice of the smoothing parameters. To understand why, let us write $R_n(\phi)$ as

$$\begin{aligned} R_n(\phi) &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \phi(X_i, \hat{P}_{ni}) + \frac{1}{n}\sum_{i=1}^{n} \Big(E(Y_i|X_i) - \hat{E}_n(Y_i|X_i) \Big) \phi(X_i, \hat{P}_{ni}) \\ &\equiv A_n(\phi) + B_n(\phi). \end{aligned} \tag{13}$$

To control the bias term $B_n(\phi)$, existing kernel methods estimate $\hat{Y}_{ni} = \hat{E}_n(Y_i|X_i)$ with a bandwidth parameter sequence h that satisfies the condition $nh^4 \to 0$ as $n \to \infty$; see DG (p. 1475). In contrast, we only require $h \to 0$ as $n \to \infty$, and standard regularity conditions on first steps. Therefore, undersmoothing is not necessary and optimal bandwidths can be used in estimating $E(Y_i|X_i)$ Moreover, our choice of ϕ makes the bias term $B_n(\phi)$ second order, while in DG it is of first order. This results in smaller bias for our test statistic relative to that of DG. Finally, in our test estimation of $E(Y_i|X_i)$ does not contribute to the first order asymptotics; this is likely to make it less sensitive to the choice of h.

The implementation of our tests, both parametric and nonparametric, depends on the specifications considered for $E(Y|X)$ and $P(X,Z)$, and on the corresponding estimators. As usual, parametric models are simplest; but they are not robust to misspecifications and may therefore invalidate our tests. This is a more serious issue for $E(Y|X)$ than for $P(X,Z)$, as estimation of $P(X,Z)$ has no impact on the asymptotic null distribution of tests (although of course it might affect power). We will consider a generic estimator $\hat{P}_{ni} \equiv \hat{P}_n(X_i, Z_i)$, which can be obtained from parametric, semiparametric or nonparametric fits in a first step. The general theory of the Appendix allows for all these possibilities. For completeness, we discuss parametric, semiparametric and nonparametric specifications for $E(Y|X)$ under the same unified theory.

3.3 Parametric inference

First consider the parametric case where

$$E(Y|X,P) = \varphi(\beta' X + \gamma P),$$

for a known real-valued function φ, an unknown vector β in a subset B of $\mathbb{R}^d$, and an unknown coefficient γ. This setting includes the standard linear regression model, with $\varphi(u) = u$, as well as probit and logit specifications. Testing for moral hazard in this setting corresponds to the parametric testing problem

$$H_0 : \gamma = 0 \qquad vs \qquad H_1 : \gamma \neq 0.$$

A standard t-test would be valid (even when P is estimated). Here we choose LM tests: they only require smoothing estimation under the null, which is convenient in later sections given the

curse of dimensionality of nonparametric estimation. An LM test statistic based on least squares residuals leads to (10), with

$$\hat{Y}_{ni} = \varphi(\hat{\beta}' X_i),\ \hat{\varepsilon}_{ni} = Y_i - \hat{Y}_{ni} \text{ and } \phi(X_i, \hat{P}_{ni}) = \dot{\varphi}(\hat{\beta}' X_i)\hat{P}_{ni},$$

where $\hat{\beta}$ is a restricted least squares estimator on the parameter space $B \subset \mathbb{R}^d$, i.e.

$$\hat{\beta} \in \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \varphi(\beta' X_i))^2$$

and $\dot{\varphi}(u) := \partial\varphi(u)/\partial u$.

We now explain how we can construct test statistics that (i) can be implemented by least squares methods; and (ii) have standard chi-square limit distributions. Applying the Mean Value Theorem to the first order condition of the nonlinear least squares objective function w.r.t. β, we can show that

$$\hat{\beta} - \beta = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(E\left(\dot{\varphi}^2(\beta' X_i) X_i X_i' \right) \right)^{-1} \dot{\varphi}(\beta' X_i) X_i + o_P(n^{-1/2}).$$

Therefore the expansion (13) holds with a bias term $B_n(\phi)$ under H_0 given by

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \left(\varphi(\beta' X_i) - \varphi(\hat{\beta}' X_i) \right) \phi(X_i, \hat{P}_{ni}) &= -\left(\hat{\beta} - \beta \right) \frac{1}{n} \sum_{i=1}^{n} \phi(X_i, \hat{P}_{ni}) \dot{\varphi}(\beta' X_i) X_i + o_P(n^{-1/2}) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (L\phi)(X_i) + o_P(n^{-1/2}), \end{aligned}$$

where $\phi(X, P)$ here is a measurable function satisfying some regularity conditions and

$$(L\phi)(X_i) = E\left(\phi(X_i, P_i) \dot{\varphi}(\beta' X_i) X_i \right) \left(E\left(\dot{\varphi}^2(\beta' X_i) X_i X_i' \right) \right)^{-1} \dot{\varphi}(\beta' X_i) X_i. \tag{14}$$

We will choose test functions ϕ so that $L\phi \equiv 0$, and hence for which the bias term becomes of second order $B_n(\phi) = o_P(n^{-1/2})$. To implement a parametric test based on S_n in (10), we recommend the following procedure:

Algorithm 3.1 *Algorithm for parametric tests:*

1. *Obtain first-step estimators* $\left\{ \hat{P}_{ni} \right\}_{i=1}^{n}$ *(parametric, semiparametric or nonparametric).*
2. *Compute restricted least squares residuals* $\left\{ \hat{\varepsilon}_{ni} = Y_i - \varphi(\hat{\beta}' X_i) \right\}_{i=1}^{n}$.
3. *For each component of an initial m-dimensional vector of test functions* ϕ_0*, compute the least squares residuals of* $\phi_0(X_i, P_i)$ *on* $\dot{\varphi}(\hat{\beta}' X_i) X_i$*, denoted by* $\phi(X_i, P_i)$*. This vector of residuals consists of new test functions that satisfy the sample analog of* $L\phi \equiv 0$.

4. *Run a regression of a vector of ones on* $\hat{\varepsilon}_{ni}\phi(X_i, P_i)$*, and compute* $S_n = nR_u^2$*, where* R_u^2 *is the uncentered* R^2 *of the regression.*

5. *Reject* H_0 *if* $S_n > \chi^2_{m,1-\alpha}$ *(the* $1-\alpha$ *quantile of the* χ^2_m *distribution).*

Wooldridge (1990) contains further discussion on the validity of this algorithm when the $\hat{P}_{ni}$ are not estimates. In the Appendix we prove the validity of this procedure in a more general setting that includes nonparametric residuals. Nonparametric tests with a continuum of weights ϕ_s can also be implemented when a parametric model is used for $E(Y|X,P)$; but to save space, we only present these nonparametric tests below in the context of nonparametric fits, i.e. when $E(Y|X)$ is nonparametric.

3.4 Semiparametric inference

A natural extension of the previous setting considers an unknown link function φ. This leads to a semiparametric single-index fit that can be obtained, for example, from the semiparametric least squares estimator of Ichimura (1993):

$$\hat{\beta} = \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - Y_{ni}(\beta)\right)^2 1(X_i \in \mathcal{X}),$$

where

$$Y_{ni}(\beta) = \frac{\sum_{j=1}^{n} Y_j k_h\left(\beta' X_i - \beta' X_j\right)}{\sum_{j=1}^{n} k_h\left(\beta' X_i - \beta' X_j\right)},$$

$k_h(u) = h^{-1}k\left(u/h\right)$, $k\left(\cdot\right)$ is a kernel function, h denotes a bandwidth parameter, and $\mathcal{X}$ is a compact and convex (non-empty) subset of $\mathbb{R}^d$ that is introduced to avoid a small random denominator in $Y_{ni}(\beta)$. Then, with $\hat{Y}_{ni} = Y_{ni}(\hat{\beta})$ we compute residuals

$$\hat{\varepsilon}_{ni} = Y_i - \hat{Y}_{ni}.$$

Given the semiparametric single index restriction, the bias term in the expansion (13) equals

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(E(Y_i|X_i) - \hat{E}_n(Y_i|X_i)\right)\phi(X_i,\hat{P}_{ni}) &= \frac{1}{n}\sum_{i=1}^{n}\left(E(Y_i|\beta' X_i) - \hat{E}_n(Y_i|\hat{\beta}' X_i)\right)\phi(X_i,\hat{P}_{ni}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E(Y_i|\beta' X_i) - \hat{E}_n(Y_i|\beta' X_i)\right)\phi(X_i,\hat{P}_{ni}) \qquad (15) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}_n(Y_i|\beta' X_i) - \hat{E}_n(Y_i|\hat{\beta}' X_i)\right)\phi(X_i,\hat{P}_{ni}). \qquad (16)
\end{aligned}$$

Lemma 5.6 in Ichimura (1993, pp. 95) shows that under regularity conditions, (16) is $o_P(n^{-1/2})$; and Theorem 3.2 in Escanciano, Jacho-Chavez and Lewbel (2014) implies that (15) equals

$$-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(L\phi)(X_i)+o_P(n^{-1/2}),$$

where

$$(L\phi)(X_i)=E\left(\phi(X_i,P_i)|\beta'X_i\right). \tag{17}$$

The following procedure implements a parametric test in the semiparametric setting:

Algorithm 3.2 *Algorithm for semiparametric tests: replace step 3 in Algorithm 3.1 with*

3': For each component of an initial m-dimensional vector of test functions ϕ_0, stimate nonparametrically the least squares residuals $\phi(X_i,P_i)=\phi_0(X_i,P_i)-\hat{E}_n\left(\phi_0(X_i,P_i)|\beta'X_i\right)$.

Again, the residuals in Algorithm 3.2 satisfy the sample analog of $L\phi\equiv 0$. As in Section 3.3, this construction allows for optimal estimators $Y_{ni}(\hat{\beta})$; it reduces the bias of the test and is less sensitive to smoothing parameters[8].

3.5 Nonparametric inference

Semiparametric methods are still subject to misspecification errors. This motivates a fully nonparametric procedure where $E(Y|X)$ is completely unspecified. Here we compute nonparametric residuals with the Nadaraya-Watson estimator

$$\hat{\varepsilon}_{ni}=Y_i-\hat{Y}_{ni}, \tag{18}$$

where

$$\begin{aligned}\hat{Y}_{ni} &= \frac{T_n(X_i)}{f_n(X_i)},\\ T_n(x) &= \frac{1}{n}\sum_{j=1}^{n}Y_jK_h\left(x-X_j\right),\\ f_n(x) &= \frac{1}{n}\sum_{j=1}^{n}K_h\left(x-X_j\right),\end{aligned} \tag{19}$$

with $K_h\left(x-X_j\right)=\prod_{l=1}^{d}k_h(x_l-X_{lj})$, and k_h is a rescaled kernel with bandwidth h.

We show in the Appendix that under regularity conditions, the following expansion holds in the nonparametric case for suitable test functions ϕ:

$$\frac{1}{n}\sum_{i=1}^{n}\left(E(Y_i|X_i)-\hat{E}_n(Y_i|X_i)\right)\phi(X_i,\hat{P}_{ni})=-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(L\phi)(X_i)+o_P(n^{-1/2}),$$

[8]Related nonparametric tests under index restrictions have been studied by Song (2009); he obtained asymptotically pivotal statistics, but for known P_i.

with

$$(L\phi)(X_i) = E\left(\phi(X_i, P_i)|X_i\right). \tag{20}$$

We will base our nonparametric tests on the following class of test functions indexed by $s = (x, p) \in \mathcal{S} \equiv \mathcal{X} \times [0, 1]$:

$$\phi_s(X_i, \hat{P}_{ni}) := 1(X_i \leq x) f_n(X_i) \left(f_n(X_i) \exp\left(p\hat{P}_{ni}\right) - \hat{r}_p(X_i) \right),$$

where

$$\hat{r}_p(x) := \frac{1}{n} \sum_{j=1}^{n} \exp\left(p\hat{P}_{nj}\right) K_h\left(x - X_j\right)$$

estimates consistently $r_p(x) := f(x)E\left(\exp(pP_i)|X_i = x\right)$ and $f_n(\cdot)$ given in (19) is a kernel estimator of the true density $f(\cdot)$ of X.

Under suitable regularity conditions, $\phi_s(X_i, \hat{P}_{ni})$ converges in mean square to

$$\phi_s^0(X_i, P_i) = 1(X_i \leq x) f^2(X_i) \left(\exp(pP_i) - E\left(\exp(pP)|X = X_i\right)\right).$$

Our nonparametric test of conditional independence is similar in nature to the nonparametric significance tests proposed in DG, which used the moments

$$\phi_s^{DG}(X_i, P_i) = 1(X_i \leq x) f_n(X_i) 1(P_i \leq p).$$

There are three important differences with respect to DG:

- We consider moment functions satisfying $L\phi = 0$, which leads to the advantages mentioned above of having small bias, avoiding undersmoothing, permitting commonly used cross-validated bandwidths and having better finite sample performance.
- In DG's setting P_i was observed, whereas in our setting we have a regressor that is nonparametrically generated. This significantly complicates inference. For example, we would require stringent conditions on the rate of convergence of the estimator $\hat{P}_{ni}$ to ensure the convergence of $1(\hat{P}_{ni} \leq p)$ [9]. This motivates our choice of the smooth test function $\exp(pP)$.
- The factor $f_n^2(X_i)$ in our test allows weakening strong conditions on the density (i.e. density bounded away from zero) that otherwise are needed in DG.
- Finally, our method of proof is very different from DG and allows for stochastic bandwidth choices, which are common in applied work (e.g. cross-validated bandwidths).

To study the asymptotic distribution of C_n in (12) we view this test statistic as a continuous functional of the stochastic process $R_n(\cdot)$ in (11). To derive its asymptotic distribution under the null hypothesis we first show that $n^{1/2}R_n$ converges weakly to a process in $\ell^\infty(\mathcal{S})$ defined below;

[9] This is similar to the difficulties with the maximum score estimator of Manski (1975).

then we apply the Continuous Mapping Theorem (CMT) (e.g. Dudley 1999, Theorem 3.6.7, pp. 116). Our next result provides the limiting distribution of $n^{1/2}R_n$ under the null. Let $\ell^\infty(\mathcal{S})$ be the Banach space of uniformly bounded functions on $\mathcal{S}$ endowed with the supremum norm, $\|f\|_\infty = \sup_{s\in\mathcal{S}} |f(s)|$. Let R_∞ be zero mean Gaussian process with covariance function,

$$K(s_1, s_2) = E(\varepsilon_i^2 \phi_{s_1}^0(X_i, P_i)\, \phi_{s_2}^0(X_i, P_i))$$

for fixed s_1 and s_2 in $\mathcal{S}$.

Theorem 3.1 *Under H_0, if A1 to A5 in the Appendix hold, then*

$$n^{1/2}R_n \text{ converges in distribution to } R_\infty \text{ in } \ell^\infty(\mathcal{S}).$$

As a consequence of Theorem 3.1 and the CMT, under H_0, p

$$C_n \to_d \int_{\mathcal{S}} \left(\frac{R_\infty(s)}{\sigma}\right)^2 d\mu(s)$$

where $\sigma^2 = E\varepsilon^2$. Unfortunately, the limiting distribution R_∞ is not pivotal, and the asymptotic critical values of C_n are difficult to compute except in special circumstances. Hence, we propose to implement the test with the assistance of a simple bootstrap procedure. We prove below that the distribution of R_∞ can be approximated by the limiting bootstrap distribution of

$$\sqrt{n}R_n^*(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\varepsilon}_{ni} \phi_s(X_i, \hat{P}_{ni}) V_i,$$

where $\{V_i, i = 1, ..., n\}$ are random variables such that $\{V_i\}_{i=1}^n$ are bounded, *iid* independent of $\mathcal{W}_n = \{Y_i, D_i, X_i, Z_i, i = 1, ..., n\}$, such that $E(V_1) = 0$ and $E(V_1^2) = 1$. Examples are Rademacher variables $\Pr(V_i = -1 | \{W_i\}_{i=1}^n) = \Pr(V_i = 1 | \{W_i\}_{i=1}^n) = 0.5$.

Let C_n^* denote the corresponding bootstrap test statistic when $\sqrt{n}R_n$ is replaced by R_n^*. As usual, we approximate the bootstrap distributions by Monte Carlo. Let $\{C_{n,b}^*\}_{b=1}^B$ be B independent Monte Carlo simulations of C_n^*, and let $c_{n,\alpha}^*$ be the corresponding α-th empirical critical value, $\alpha \in (0, 1)$. We reject the null hypothesis of moral hazard with the nonparametric test at the α-th significance level if $C_n > c_{n,\alpha}^*$. More generally, the unknown limiting null distribution of $g(R_n)$, i.e. the distribution of $g(R_\infty)$, is approximated by the bootstrap distribution of $g(R_n^*)$. That is, the bootstrap distribution

$$F^*_{g(\sqrt{n}R_n^*)}(x) = \Pr(g(R_n^*) \le x | \{W_i\}_{i=1}^n)$$

estimates the asymptotic null distribution function

$$F_{g(R_\infty)}(x) = \Pr(g(R_\infty) \le x).$$

Thus, H_0 will be rejected at the $100\alpha\%$ of significance when $g(R_n) \ge c_{n,\alpha}^*$, where $F^*_{g(R_n^*)}(c_{n,\alpha}^*) = 1 - \alpha$. The bootstrap assisted test is valid if $F^*_{g(R_n^*)}$ is a consistent estimator of $F_{g(R_\infty)}$ at each continuity point of $F_{\varphi(R_\infty)}$. When consistency is in probability it is expressed as $g(R_n^*) \to_d g(R_\infty)$ *in probability*. See Giné and Zinn (1990) or van der Vaart and Wellner (1996) for discussion.

In the Appendix we prove the validity of the bootstrap approximations.

Theorem 3.2 *If A*1 *to A*5 *in the Appendix hold, for any continuous functional g in* $\ell^{\infty}(\mathcal{S})$,

$$g(R_n^*) \to_d g(R_\infty) \text{ in probability.}$$

Remark 3.2 *Our choice of functional* $g(R) = \int_{\mathcal{S}} \left(\frac{R(s)}{\sigma}\right)^2 d\mu(s)$ *involves unknown quantities in* σ *and perhaps* μ. *However, by standard arguments, estimation of these objects does not affect the asymptotic null distribution of our test.*

We recommend the following procedure to implement the nonparametric test.

Algorithm 3.3 *Algorithm for nonparametric tests:*

1. *Obtain first-step estimators* $\left\{\hat{P}_{ni}\right\}_{i=1}^{n}$ *(parametric, semiparametric or nonparametric).*
2. *Compute nonparametric restricted residuals as in (18) and test statistic* C_n *as in the closed form expression (21) in page 23.*
3. *Generate n iid Rademacher variables* $\{V_i\}_{i=1}^n$, *i.e.* $\Pr(V_i = -1) = \Pr(V_i = 1) = 0.5$, *independent of the original sample. Compute bootstrap residuals* $\hat{\varepsilon}_{ni}^* = \hat{\varepsilon}_{ni} V_i$ *and compute* C_n^* *as* C_n *in (21) in page 23 but with* $\hat{\varepsilon}_{ni}$ *replaced by* $\hat{\varepsilon}_{ni}^*$.
4. *Repeat Step 3 B times, and compute the* $(1-\alpha)-$*empirical quantile of the obtained bootstrap test statistics,* $C_{n,j}^*$ $j = 1, ..., B$, *say* $c_{n,\alpha}^*$.
5. *Reject* H_0 *at level* α *if* $C_n > c_{n,\alpha}^*$.

4 Testing for stochastic dominance

Proposition 2.1 established that in the absence of moral hazard, outcomes Y and instruments Z are independent conditional on $X = x$. But sometimes the economic structure of the problem indicates the direction that the codependence of Y and Z will take under moral hazard. In an insurance context for instance, effort is often assumed to shift the mean of the distribution of losses to the left, or perhaps only to generate a distributions of losses that are ordered by first-order stochastic dominance. If in addition contract choice is monotonic with respect to instruments Z, then intuition suggests that a higher value of the instrument should shift the distribution of losses to the left.

To make this more precise, we return to the setting of the model of section 2. Let z and z' be two values of the instruments Z, and x be a value of the covariates. The following gives sufficient conditions for the distributions of outcomes to be ordered by the values of the instruments:

Assumption 4.1 (Stochastic Dominance at (x, z, z')**)**

(i) there exists an ordering of contract choices $\succ_D$ for which the effort function $(D, V) \to e(x, D, V)$ is strictly increasing with D for all V

(ii) for the same ordering $\succ_D$, $h(x, z', V) \succ_D h(x, z, V)$ for all V

(iii) $(e, V, \eta) \to g(e, x, V, \eta)$ is increasing in e.

The combination of (i) and (ii) in Assumption 4.1 implies that when instruments shift from z to z', the distribution of effort shifts to the right. Part (iii) then implies that so does the distribution of outcomes. Obviously, part (i) of the assumption violates the null hypothesis of no moral hazard.

In an insurance model, the order $\succ_D$ in the assumption would simply reflect lower coverage, which induces higher effort. The change from z to z' could come from a higher price of coverage, or an insurer offering a new contract with lower coverage.

Proposition 4.1 *Under Assumptions 2.1 and 4.1, the distribution of Y conditional on $(Z = z', X = x)$ first-order stochastically dominates that of Y conditional on $(Z = z, X = x)$.*

Proof: remember that

$$Y = g\left(e(D, x, V), x, V, \eta\right) = g\left(e(h(x, Z, V), x, V), x, V, \eta\right).$$

Therefore for any y, using Assumption 4.1.(iii),

$$\Pr(Y < y | Z = z, X = x) = \Pr(e(h(x, z, V), x, V) < \bar{g}(y, x, V, \eta) | X = x)$$

for an increasing function $\bar{g}$. Using part (i),

$$\Pr(Y < y | Z = z, X = x) = \Pr(h(x, z, V) \prec_D \tilde{g}(y, x, V, \eta) | X = x)$$

for some function $\tilde{g}$, and finally, part (ii) gives

$$\Pr(h(x, z, V) \prec_D \tilde{g}(y, x, V, \eta) | X = x) > \Pr(h(x, z', V) \prec_D \tilde{g}(y, x, V, \eta) | X = x)$$

so that $\Pr(Y < y | Z = z, X = x) > \Pr(Y < y | Z = z', X = x)$. □

Proposition 4.1 suggests the use of tests that focus on the alternative of stochastic dominance. In a parametric setting, similar to that of Section 3.3, developing such methods is straightforward. For example, consider for simplicity the binary outcome case and a probit or logit type specification

$$E(Y | X, Z) = \varphi(\beta' X + \gamma Z).$$

The null of conditional independence corresponds to $\gamma = 0$. The alternative of stochastic dominance corresponds to $\gamma > 0$ or $\gamma < 0$, depending on the direction of monotonicity. A simple one sided t-test would be a valid test. This can be implemented in off-the-shelf statistical software.

Nonparametric tests of stochastic monotonicity are also available in the literature. See for example the test proposed by Romano, Shaikh and Wolf (2014), which can be applied when the support of the variables (Y, X, Z) is discrete, or Lee, Linton and Whang (2009), Delgado and Escanciano (2012), and Andrews and Shi (2013), which can be applied to continuous variables.

Concluding remarks

While we chose to only discuss inference in detail for the 2-by-2 model, it should be clear from section 2 that our approach is much more general. Applying it requires checking that Assumption 2.2 holds, so that the instruments have power to detect deviations from the null, and then writing the appropriate test statistic. Both steps will be slightly different depending on the range of variation of instruments, of efforts, and of outcomes; but the same principles apply. In the 2-by-2 model we recommend first checking the relevance of instruments by applying the same conditional independence test proposed above but where Y_i and P_i are replaced, respectively, by D_i and Z_i. Then, if the instruments are found nonparametrically relevant, proceed by applying the nonparametric test based C_n, with the closed form expression provided in the Appendix.

The contribution of this approach to testing for moral hazard depends on its power to reject the null in the kind of samples that researchers use in practice. We are currently developing a Monte Carlo simulation study to test how our procedure performs in realistic instances of the 2-by-2 model.

References

Andrews, D.W.K. and X. Shi (2013), "Inference based on conditional moment inequalities," *Econometrica*, 81, 609-666.

Annan, F. (2015), "Identifying and Estimating Asymmetric Information using an Instrument from a National Reform", mimeo Columbia.

Chiappori, P.-A., F. Durand, and P. Y. Geoffard, 1998,, "Moral Hazard and the Demand for Physician Services: First Lessons From a French Natural Experiment", in European Economic Review, 42, 499–511.

Chiappori, P.-A., B. Jullien, B. Salanié and F. Salanié (2006), "Asymmetric information in insurance: general testable implications", *Rand Journal of Economics*, 35, 783–798.

Chiappori, P.-A. and B. Salanié (2014), "Asymmetric Information in Insurance Markets: Predictions and Tests", in *Handbook of Insurance*, 2nd edition (G. Dionne, ed).

Cohen, A. and L. Einav (2007), "Estimating Risk Preferences from Deductible Choice", *American Economic Review*, 97, 745–788.

Delgado, M.A., Escanciano, J.C. (2012): "Distribution-free tests of stochastic monotonicity," *Journal of Econometrics* 170, 68–75.

Delgado, M.A. and Gonzalez-Manteiga, W. (2001): "Significance testing in nonparametric regression based on the bootstrap," *Annals of Statistics*, 29, 1469-1507.

Dionne, G. and M. Vanasse (1997), ""Une évaluation empirique de la nouvelle tarification de l'assurance automobile au Québec", *L'actualité économique*, 73, 47—80.

Dudley, R.M. (1999). *Uniform Central Limit Theorems.* Cambridge. University Press, Massachusetts.

Einav L., Finkelstein A., Ryan S., Schrimpf P., and M.-R. Cullen (2013): "Selection on Moral Hazard in Health Insurance," *American Economic Review*, 103, 178–219.

Escanciano, J.C., Jacho-Chavez, D. and Lewbel, A. (2014): "Uniform convergence of weighted sums of non- and semi-parametric residuals for estimation and testing," *Journal of Econometrics*, 178, 426-443.

Giné, E. and Zinn, J. (1990): "Bootstrapping general empirical measures," *Annals of Probability,* 18, 851-869.

Handel, B. (2013): "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurt", *American Economic Review*, 103, 2643–2682.

Heckman, J., and R. Pinto (2015), "Unordered monotonicity," University of Chicago, mimeo.

Heckman, J., S. Urzua, and E. Vytlacil (2006), "Understanding instrumental variables in models with essential heterogeneity," *Review of Economics and Statistics*, 88, 389—432.

Heckman, J., S. Urzua, and E. Vytlacil (2008), "Instrumental variables in models with multiple outcomes: The general unordered case," *Annales d'economie et de statistique*, 151—174.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58(1-2), 71–120.

Lee, S., O. Linton, and Y. Whang. (2009): "Testing for Stochastic Monotonicity," *Econometrica,* 77, 585-602.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987), "Health Insurance and the Demand for Medical Care: Evidence From a Randomized Experiment", *American Economic Review*, 77, 251–277.

Romano, J., A. Shaikh, and M. Wolf (2014), "A practical two-step method for testing moment inequality models," *Econometrica*, 82, 1979-2002.

Salanié, B. (2005), *The economics of contracts: a primer*, 2nd ed, MIT Press.

Song, K. (2009), "Testing Conditional Independence via Rosenblatt Transforms," *Annals of Statistics* 37, 4011–4045.

Su, L. and White, H. (2007), "A consistent characteristic function-based test for conditional independence," *Journal of Econometrics*, 141, 807–34.

Su, L. and White, H. (2008), "A nonparametric Hellinger metric test for conditional independence," *Econometric Theory*, 24, 829–64.

van der Vaart, A. W. (1998), *Asymptotic statistics*, vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

van der Vaart, A.W. and Wellner, J.A (1996). *Weak converngence and empirical processes*. Springer. New York.

van der Vaart, A.W. and Wellner, J.A. (2007), "Empirical processes indexed by estimated functions," in *Asymptotics: Particles, Processes and Inverse Problems*. Institute of Mathematical Statistics Lecture Notes—Monograph Series 55 234–252. Beachwood, OH: IMS.

Weisburd, S. (2013), "Identifying Moral Hazard in Car Insurance Contracts", mimeo Tel Aviv.

Wooldridge, J.M. (1990), "A unified approach to robust, regression-based specification tests", *Econometric Theory,* 6, 17-43.

5 Appendix:

5.1 Computation of the nonparametric test statistic

The nonparametric test statistic (12) can be easily computed in closed form when the integrating measure is given by

$$d\mu(s) = dF_n(x) \times dp,$$

where F_n is the empirical cumulative distribution function of covariates $\{X_i\}_{i=1}^n$. With this choice, after some simple algebra C_n is the following quadratic form of residuals $\{\hat{\varepsilon}_{ni}\}_{i=1}^n$

$$C_n = \frac{1}{n^2\sigma_n^2}\sum_{i=1}^n\sum_{k=1}^n \hat{\varepsilon}_{ni}\hat{\varepsilon}_{nk}a_{ik}b_{ik}, \tag{21}$$

where

$$a_{ik} = \sum_{j=1}^n 1(X_i \leq X_j)1(X_k \leq X_j),$$

$$b_{ik} = f_n(X_i)f_n(X_k)\left[f_n(X_i)f_n(X_k)\frac{\exp\left(\hat{P}_{ni}+\hat{P}_{nk}\right)-1}{\hat{P}_{ni}+\hat{P}_{nk}} - f_n(X_i)c_{ik} - f_n(X_k)c_{ki} + d_{ik}\right],$$

$$c_{ik} = \frac{1}{n}\sum_{j=1}^n\left(\frac{\exp\left(\hat{P}_{ni}+\hat{P}_{nj}\right)-1}{\hat{P}_{ni}+\hat{P}_{nj}}\right)K_h\left(X_k - X_j\right)$$

and

$$d_{ik} = \frac{1}{n^2}\sum_{j=1}^n\sum_{l=1}^n\left(\frac{\exp\left(\hat{P}_{nj}+\hat{P}_{nl}\right)-1}{\hat{P}_{nj}+\hat{P}_{nl}}\right)K_h\left(X_i - X_j\right)K_h\left(X_k - X_l\right).$$

5.2 Proofs of main inference results

The sample observations $\{W_i = (Y_i, D_i, X_i, Z_i)\}_{i=1}^n$ are a sequence of independent and identically distributed (iid) variables defined on the measurable space $(\mathcal{W}, \mathcal{B})$, with probability law $\mathbb{P}$, and distributed as $\{W = (Y, D, X, Z)\}$. Let $\mathbb{P}_n$ denote the empirical distribution associated to $\{W_i\}_{i=1}^n$. Henceforth, for a measurable function g we denote

$$\mathbb{P}g = \int g d\mathbb{P},$$

where we drop the domain of integrations to simplify the notation. The empirical process is defined as

$$\begin{aligned}\mathbb{G}_n g &= \sqrt{n}\left(\mathbb{P}_n g - \mathbb{P}g\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{g(W_i) - E\left(g(W_i)\right)\right\}.\end{aligned}$$

To measure the complexity of a class of functions $\mathcal{G}$, we introduce the concept of bracketing numbers. An envelope function G for the class $\mathcal{G}$ is a measurable function such that $G(x) \geq \sup_{g\in\mathcal{G}} |g(x)|$. Given two functions l, u, a bracket $[l, u]$ is the set of functions $f \in \mathcal{G}$ such that $l \leq f \leq u$. An ε-bracket with respect to a norm $\|\cdot\|$ is a bracket $[l, u]$ with $\|l - u\| \leq \varepsilon$, $\|l\| < \infty$ and $\|u\| < \infty$ (note that u and l not need to be in $\mathcal{G}$). The *covering number with bracketing* $N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|)$ is the minimal number of ε-brackets with respect to $\|\cdot\|$ needed to cover $\mathcal{G}$. Let $\|\cdot\|_2$ denote the L_2-norm $\|g\|_2^2 := \int g^2 d\mathbb{P}$. Define for any vector $(a_1, ..., a_d)$ of d integers the differential operator $\partial_x^a := \partial^{|a|}/\partial x_1^{a_1} \dots \partial x_d^{a_d}$, where $|a| := \sum_{i=1}^d a_i$.

Let $f(x)$ denote the density of X evaluated at a point x in its support. For any smooth function $h : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ and some $\eta > 0$, let $\underline{\eta}$ be the largest integer smaller than η, and

$$\|h\|_{\infty,\eta} := \max_{|a|\leq\underline{\eta}} \sup_{x\in\mathcal{X}} |\partial_x^a h(x)| + \max_{|a|=\underline{\eta}} \sup_{x_1\neq x_2} \frac{|\partial_x^a h(x_1) - \partial_x^a h(x_2)|}{|x_1 - x_2|^{\eta-\underline{\eta}}}.$$

Further, let $C_M^\eta(\mathcal{X})$ be the set of all continuous functions $h : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ with $\|h\|_{\infty,\eta} \leq M$. Then, it is known that $\ln N_{[\cdot]}(\varepsilon, C_M^\eta(\mathcal{X}), \|\cdot\|_\infty) \leq C\varepsilon^{-v_w}$, with $v_w = d/\eta$.

Henceforth, $\mathcal{F}$ is a class of measurable functions with $\mathcal{F} \subset C_{M_f}^{\eta_f}(\mathcal{X})$ and $\mathcal{T}$ is class of measurable functions with $\mathcal{T} \subset C_{M_T}^{\eta_T}(\mathcal{X})$. Let $\mathcal{P}$ and Ψ denote classes of measurable functions of $(X, Z) \subset \mathbb{R}^{d+p}$. Let $\hat{P}(X_i, Z_i)$ be an estimator of $P(X, Z) = E(D|X, Z)$. Define $m(x) := E(Y|X = x)$, $T(x) := f(x) \times m(x)$, and for $p \in [-1, 1]$ and $\pi \in \mathcal{P}$,

$$r_{p,\pi}(x) := f(x)E\left(\exp(p\pi(X_i, Z_i))|X_i = x\right).$$

A1.- $E[Y^2] < \infty$ and $\mathcal{X}$ is a convex, compact subset of $\mathbb{R}^d$, with non-empty interior.

A2.- (i) f and T are uniformly continuous; (ii) $T \in C_{M_T}^{\eta_T}(\mathcal{X})$, $f \in C_{M_f}^{\eta_f}(\mathcal{X})$, $\mathbb{P}(T_n \in \mathcal{T}) \to 1$, and $\mathbb{P}(f_n \in \mathcal{F}) \to 1$, for some $\eta_T > d/2$, $\eta_f > d/2$ and $M_f < \infty$.

A3.- (i) $r_{p,\pi}(\cdot)$ is uniformly equicontinuous, i.e.

$$\lim_{\delta\to 0} \sup_{x,|z|<\delta} \sup_{p\in[-1,1],\pi\in\mathcal{P}} |r_{p,\pi}(x - z) - r_{p,\pi}(x)| = 0;$$

$\left\|\hat{P} - P\right\|_2 = o_P(1)$, $P \in \mathcal{P}$ and $\mathbb{P}(\hat{P} \in \mathcal{P}) \to 1$, where $\mathcal{P}$ is a class of measurable functions from $\mathbb{R}^{d+p}$ to $[0, 1]$; (ii) $\ln N_{[\cdot]}(\varepsilon, \mathcal{P}, \|\cdot\|_2) \leq C\varepsilon^{-v_P}$ for some $v_P < 2$. Furthermore, $\mathbb{P}(\hat{r}_p \in \mathcal{F}) \to 1$ for each $p \in [-1, 1]$.

A4.- The kernel function $k(t) : \mathbb{R} \to \mathbb{R}$ is symmetric, bounded, integrable and satisfies the following conditions: $\int k(t)\, dt = 1$, $\int k^2(u)du < \infty$ and $\lim_{|z|\to\infty} |z|\, |k(z)| = 0$.

A5.- The bandwidth $h \equiv h_n$ satisfies $\Pr(a_n \leq h_n \leq b_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers a_n and b_n such that $b_n \to 0$ and $a_n^d n/\log n \to \infty$.

Assumption 1 is standard. The compact support of X can be relaxed. Assumptions A2(ii)-A3 follow from certain smoothness in the underlying class of functions. Assumption A4 is standard in the literature of nonparametric kernel estimation, while Assumption A5 permits data-dependent bandwidths.

Proof of Theorem 3.1: Define

$$\hat{\psi}_s(X_i, \hat{P}_{ni}) := 1(X_i \leq x) \left(f_n(X_i) \exp(p\hat{P}_{ni}) - \hat{r}_p(X_i) \right)$$

and then write

$$\begin{aligned} R_n(s) &= \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_{ni} \phi_s(X_i, \hat{P}_{ni}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[Y_i f_n(X_i) - T_n(X_i) \right] \hat{\psi}_s(X_i, \hat{P}_{ni}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_i f(X_i) - T_n(X_i) \right) \hat{\psi}_s(X_i, \hat{P}_{ni}) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} Y_i \left(f_n(X_i) - f(X_i) \right) \hat{\psi}_s(X_i, \hat{P}_{ni}) \\ &\equiv R_{1n}(s) + R_{2n}(s). \end{aligned} \tag{22}$$

To analyze $R_{1n}(s)$ define the classes of functions

$$\Psi := \{(x,z) \to 1(\bar{x} \leq x) \left(\eta(x) \exp(p \times \pi(x,z)) - r(x) \right) : (\bar{x}, p) \in \mathcal{S}; \eta, r \in \mathcal{F}; \pi \in \mathcal{P}\}$$

and

$$\mathcal{G}_1 := \{(y,x,z) \to \left(yf(x) - \varphi(x) \right) \psi(x,z) : \varphi \in \mathcal{T}, \psi \in \Psi\}.$$

By Lemma A1 below, $\mathcal{G}_1$ is $\mathbb{P}$-Donsker if $\mathcal{T}$ and Ψ are $\mathbb{P}$-Donsker. Since $\mathcal{T} \subset C^{\eta_T}_{M_T}(\mathcal{X})$ with $\eta_T > d/2$, then it is known that $\mathcal{T}$ is $\mathbb{P}$-Donsker. Lemma A3 below shows that Ψ is $\mathbb{P}$-Donsker.

Define

$$g_{s,\eta_n}(W_i) := \left(Y_i f(X_i) - T_n(X_i) \right) \hat{\psi}_s(X_i, \hat{P}_{ni})$$

and

$$g_{s,\eta_0}(W_i) := \left(Y_i f(X_i) - T(X_i) \right) \psi_s(X_i, P_i),$$

where $\eta_n = (T_n, \hat{\psi}_s)$ and $\eta_0 = (T, \psi_s)$, with $\psi_s(X_i, P_i) := 1(X_i \leq x) \left(f(X_i) \exp(pP_i) - r_p(X_i) \right)$.

Now, using that ψ_s and T are uniformly bounded and the triangle and Cauchy-Schwarz inequalities, write

$$\begin{aligned} \sup_{s \in \mathcal{S}} \left\| g_{s,\eta_n} - g_{s,\eta_0} \right\|_2 &\leq C \left\| T_n - T \right\|_2 + C \sup_{s \in \mathcal{S}} \left\| \hat{\psi}_s - \psi_s \right\|_2 + \left\| T_n - T \right\|_2 \sup_{s \in \mathcal{S}} \left\| \hat{\psi}_s - \psi_s \right\|_2 \\ &= o_P(1), \end{aligned}$$

where the last equality uses Lemmas A5-A6 below. In addition, Assumptions A2-A3, yield $T \in C^{\eta_T}_{M_T}(\mathcal{X})$, $\psi_s \in \Psi$, $\mathbb{P}(T_n \in \mathcal{T}) \to 1$ and $\mathbb{P}(\hat{\psi}_s \in \Psi) \to 1$. Conclude from Lemma A2 that uniformly in $s \in \mathcal{S}$,

$$\begin{aligned}\sqrt{n}R_{1n}(s) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{(Y_i f(X_i) - T(X_i))\,\psi_s(X_i,P_i)\} \\ &\quad +\sqrt{n}\mathbb{P}(\,(Y_i f(X_i) - T_n(X_i))\,\hat{\psi}_s(X_i,\hat{P}_{ni})) \\ &\quad +o_P(1), \\ &\equiv A_{1n}(s) + B_{1n}(s) \end{aligned} \tag{23}$$

We first analyze the bias term $B_{1n}(s)$. By the triangle inequality, Cauchy-Schwarz inequality and the moment

$$E\left(\psi_s(X_i,P_i)|X_i\right) = 0 \text{ a.s.}, \tag{24}$$

we have, uniformly in $s \in \mathcal{S}$,

$$\begin{aligned}B_{1n}(s) &= \sqrt{n}\mathbb{P}(\,(T(X_i) - T_n(X_i))\,\psi_s(X_i,P_i)) + o_P(1) \\ &= o_P(1).\end{aligned}$$

Similarly as for R_{1n},

$$\begin{aligned}\sqrt{n}R_{2n}(s) &= \sqrt{n}\mathbb{P}(Y_i\,(f(X_i) - f_n(X_i))\,\hat{\psi}_s(X_i,\hat{P}_{ni})) + o_P(1), \\ &= \sqrt{n}\mathbb{P}(Y_i\,(f(X_i) - f_n(X_i))\,\psi_s(X_i,P_i)) + o_P(1), \\ &= o_P(1), \end{aligned} \tag{25}$$

where the first equality uses Lemma A2, the second equality uses the triangle inequality and Cauchy-Schwarz's inequality and the last equality uses the moment restriction (24). Then, from (22), (23) and (25) we conclude, uniformly in $s \in \mathcal{S}$ and $a_n \leq h \leq b_n$,

$$R_n(s) = \frac{1}{n}\sum_{i=1}^{n}\left\{\varepsilon_i \phi^0_s(X_i,P_i)\right\} + o_P(n^{-1/2}).$$

It is straightforward to show from our results that the class

$$\{w \to \varepsilon 1(\bar{x} \leq x)\,(f(x)\exp(p \times P(x,z)) - r_p(x)) : s \in \mathcal{S}\}$$

is $\mathbb{P}$-Donsker. This completes the proof of the Theorem. □

Proof of Theorem 3.2: Write

$$\begin{aligned}\sqrt{n}R^*_n(s) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\phi^0_s(X_i,P_i)V_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\hat{\varepsilon}_{ni}\phi_s(X_i,\hat{P}_{ni}) - \varepsilon_i\phi^0_s(X_i,P_i)\}V_i \\ &:= R^*_{1n}(s) + R^*_{2n}(s).\end{aligned}$$

First, note that Theorem 2.11.9 in van der Vaart and Wellner (1996) implies

$$R^*_{1n} \to_d R_\infty \textit{ in probability.}$$

Second, we prove that

$$R^*_{2n} \to_p 0 \textit{ in probability.}$$

For this we need to show that the finite dimensional distributions of the process R^*_{2n} (conditionally on the sample) converge to zero in probability and that R^*_{2n} is asymptotically tight. Let consider a finite set of points of $\mathcal{S}$, $s_1 = (x_1', p_1)', ..., s_r = (x_1', p_1)'$ and a real vector $\lambda = (\lambda_1, ..., \lambda_r)'$ with $|\lambda| = 1$. Define

$$Z^*_{n,r} = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{r} \lambda_j \{\hat{\varepsilon}_{ni} \phi_{s_j}(X_i, \hat{P}_{ni}) - \varepsilon_i \phi^0_{s_j}(X_i, P_i)\} V_i := \sum_{i=1}^{n} \zeta^{r*}_{ni},$$

where ζ^{r*}_{ni} is implicitly defined. Then, note that conditional on the original data, ζ^{r*}_{ni} is an independent (not identically distributed) array of random variables, with

$$E^* \left(\sum_{i=1}^{n} \zeta^{r*}_{ni} \right) = 0,$$

while

$$\begin{aligned} V^* \left(\sum_{i=1}^{n} \zeta^{r*}_{ni} \right) &= \sum_{i=1}^{n} V^* \left(\zeta^{r*}_{ni} \right) \\ &= \sum_{j=1}^{r} \sum_{h=1}^{r} \lambda_j \lambda_h \left(n^{-1} \sum_{i=1}^{n} \hat{l}_{nij} \hat{l}_{nih} \right) := \widetilde{\sigma}^2_{n,r}, \end{aligned}$$

where $\hat{l}_{nij} = \{\hat{\varepsilon}_{ni} \phi_{s_j}(X_i, \hat{P}_{ni}) - \varepsilon_i \phi^0_{s_j}(X_i, P_i)\}$. By Cauchy-Schwarz's inequality and L_2 convergence

$$\begin{aligned} \left(n^{-1} \sum_{i=1}^{n} \hat{l}_{nij} \hat{l}_{nih} \right) &\leq \left(n^{-1} \sum_{i=1}^{n} \hat{l}^2_{nij} \right) \left(n^{-1} \sum_{i=1}^{n} \hat{l}^2_{nih} \right) \\ &= o_P(1), \end{aligned}$$

which follows by Markov's inequality and from routine arguments in nonparametric kernel estimation.

Next, the almost sure asymptotic uniform equicontinuity follows from Theorem 2.11.9 in van der Vaart and Wellner (1996). □

5.3 Lemmas

The following Lemma is a well-known result in empirical processes theory (see e.g. Lemma A.1 in Escanciano et al. (2014)).

Lemma A1. *Let $\mathcal{F}$ and $\mathcal{G}$ be classes of functions with a bounded and a squared integrable envelope F and G, respectively, then*

$$N_{[\cdot]}(\epsilon, \mathcal{F}\cdot\mathcal{G}, \|\cdot\|_2) \leq N_{[\cdot]}(C\epsilon, \mathcal{F}, \|\cdot\|_2) \times N_{[\cdot]}(C\epsilon, \mathcal{G}, \|\cdot\|_2).$$

The following Lemma is Theorem 2.1 in Van der Vaart and Wellner (2007) and is stated here for exposition.

Lemma A2. *Suppose the following conditions hold*

- $\{g_{s,\eta} : s \in \mathcal{S}, \eta \in \mathcal{H}\}$ *is a $\mathbb{P}$-Donsker class of measurable functions,*
- $\mathbb{P}(\eta_n \in \mathcal{H}) \to 1$ *and* $\eta_0 \in \mathcal{H}$, *as* $n \to \infty$, and
- $\sup_{s\in\mathcal{S}} \left\|g_{s,\eta_n} - g_{s,\eta_0}\right\|_2 \to 0$, *as* $n \to \infty$.

Then, uniformly in $s \in \mathcal{S}$,

$$\mathbb{G}_n g_{s,\eta_n} = \mathbb{G}_n g_{s,\eta_0} + o_P(1).$$

Lemma A3. *Ψ is a uniformly bounded $\mathbb{P}$-Donsker class.*

Proof of Lemma A3. The proof follows from Lemma A1 and the $\mathbb{P}$-Donsker property of the class

$$\{(x,z) \to \exp(p \times \pi(x,z)) : p \in [-1,1], \pi \in \mathcal{P}\},$$

which in turn follows easily from the Lipschitz property of $\pi \to \exp(p \times \pi(x,z))$. □

Lemma A4 (Uniform in bandwidth bias). Let $K : \mathbb{R}^d \to \mathbb{R}$ be a measurable bounded function such that

$$\int_{\mathbb{R}^d} |K(z)|\, dz < \infty \text{ and } \lim_{|z|\to\infty} |z|^d |K(z)| = 0.$$

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous function such that

$$\int_{\mathbb{R}^d} |g(z)|\, dz < \infty.$$

Then,

$$\sup_{a_n \leq h \leq b_n} \left| \int \frac{1}{h^d} K\left(\frac{z}{h}\right) g(x-z) dz - g(x) \int K(z)\, dz \right| = o(1). \tag{26}$$

Furthermore, if g is uniformly continuous, the convergence in (26) is uniform in $\mathbb{R}^d$.

Proof of Lemma A4. Take any $\delta > 0$. It is easy to check that the left hand side of (26) is bounded by

$$\begin{aligned}
&\sup_{|z|\leq\delta} |g(x-z)-g(x)| \int_{\mathbb{R}^d} |K(z)|\, dz \\
&+ \sup_{|z|>\delta b_n^{-1}} |K(z)|\, |z|^d\, \delta^{-d} \int_{\mathbb{R}^d} |g(z)|\, dz \\
&+ |g(x)| \int_{|z|>\delta b_n^{-1}} |K(z)|\, dz.
\end{aligned}$$

Continuity of g and $b_n \to 0$ implies the result. Uniformity in $\mathbb{R}^d$ follows from g being bounded uniformly. □

Lemma A5. Under Assumptions A1, A2(i), A4(i) and A5(i),

$$\sup_{a_n\leq h\leq b_n} \|f_n - f\|_2 = o_P(1) \text{ and } \sup_{a_n\leq h\leq b_n} \|T_n - T\|_2 = o_P(1).$$

Proof of Lemma A5. We prove $\|T_n - T\|_2 = o_P(1)$; we omite the proof of $\|f_n - f\|_2 = o_P(1)$, which is similar but simpler.

It suffices to prove that

$$\sup_{a_n\leq h\leq b_n} \int \left(E\left(T_n(x)\right) - T(x)\right)^2 f(x)dx = o(1) \tag{27a}$$

and

$$\sup_{a_n\leq h\leq b_n} \int Var\left(T_n(x)\right) f(x)dx = o(1). \tag{28}$$

The proof of (27a) follows from the proof of Lemma A4, by the uniform continuity of T and the boundedness of f. Moreover, by well known results using Assumption A1 and the boundedness of f,

$$\begin{aligned}
\sup_{a_n\leq h\leq b_n} \int Var\left(T_n(x)\right) f(x)dx &\leq \frac{1}{a_n^d n} \int\int E(Y_j^2 K_h^2\left(x - X_j\right)) f(x)dx \\
&\leq \frac{C}{a_n^d n} \int k^2(u)du \\
&= o(1).
\end{aligned}$$

This proves the Lemma. □

Lemma A6. Under Assumptions A1,A2(i), A3(i), A4(i) and A5(i),

$$\sup_{a_n\leq h\leq b_n} \sup_s \left\| \hat{\psi}_s - \psi_s \right\|_2 = o_P(1).$$

Proof of Lemma A6. For $p \in [-1,1]$ and $\pi \in \mathcal{P}$, we define

$$\hat{r}_{p,\pi}(x) := \frac{1}{n}\sum_{j=1}^{n} \exp\left(p\pi(X_j, Z_j)\right) K_h\left(x - X_j\right).$$

We first show that

$$\sup_{a_n \le h \le b_n} \sup_{p\in[-1,1]} \sup_{\pi\in\mathcal{P}} \left\|\hat{r}_{p,\pi} - r_{p,\pi}\right\|_2 = o_P(1).$$

This follows from the same arguments as Lemma A5; the bias is uniformly negligible by the uniform equicontinuity and the uniform integrability of $r_{p,\pi}$, since $\exp\left(p\pi(X_j, Z_j)\right)$ is uniformly bounded.

The uniform boundedness and Lipschitz property of the function $\exp\left(p\pi(X_j, Z_j)\right)$ also imply, uniformly in $a_n \le h \le b_n$,

$$\begin{aligned}
\sup_s \left\|\hat{\psi}_s - \psi_s\right\|_2 &\le C\left[\left\|f_n - f\right\|_2 + \left\|\hat{P} - P\right\|_2 + \sup_p \left\|\hat{r}_p - r_p\right\|_2\right] \\
&= o_P(1),
\end{aligned}$$

by Lemma A5, Assumption A3 and

$$\begin{aligned}
\sup_p \left\|\hat{r}_p - r_p\right\|_2 &\le \sup_{p\in[-1,1]} \sup_{\pi\in\mathcal{P}} \left\|\hat{r}_{p,\pi} - r_{p,\pi}\right\|_2 + \sup_{p\in[-1,1]} \left\|r_{p,\hat{P}} - r_p\right\|_2 \\
&\le o_P(1) + C\left\|\hat{P} - P\right\|_2 \\
&= o_P(1).
\end{aligned}$$

This proves the Lemma. □