Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables¹
(Preliminary and Incomplete)

Joseph G. Altonji
Timothy Conley
Todd E. Elder
Christopher R. Taber

January 23, 2015

¹We have received helpful comments from seminar participants at a number of institutions. We also thank Don Andrews, Bruce Hansen, Joel Horowitz, Jack Porter, and Xiaoxia Shi for helpful comments. We are grateful for financial support from the National Institute of Child Health and Development grant R01 HD36480-03 (Altonji and Taber) and from the Economic Growth Center, Yale University (Altonji).
Abstract

We develop new estimation methods for estimating causal effects based on the idea that the amount of selection on the observed explanatory variables in a model provides a guide to the amount of selection on the unobservables. We discuss two approaches, one of which involves the use of a factor model as a way to infer properties of unobserved covariates from the observed covariates. We construct an interval estimator that asymptotically covers the true value of the causal effect, and we propose related confidence regions that cover the true value with fixed probability.
1 Introduction

Distinguishing between correlation and causality is the most difficult challenge faced by empirical researchers in the social sciences. Social scientists are rarely in a position to run a well controlled experiment. Consequently, they rely on a priori restrictions about the relationships between the variables that are observed or unobserved. These restrictions are typically in the form of exclusion restrictions or assumptions about the functional form of the model, the distribution of the unobserved variables, or dynamic interactions. Occasionally, the restrictions are derived from a widely accepted theory or are supported by other studies that had access to a richer set of data. However, in most cases, doubt remains about the validity of the identifying assumptions and the inferences that are based on them. This reality has lead a number of researchers to focus on the estimation of bounds under weaker assumptions than those that are conventionally imposed.

In this paper, we develop estimation strategies that are useful in cases in which doubt remains about the exogeneity of instrumental variables or the treatment itself. This is the situation in many applications in economics and the other social sciences, with examples including the effectiveness of private schools, the effects of education on crime, the effects of crime on labor market outcomes, or the effects of obesity on health outcomes. Our approach uses the degree of selection on observed variables as a guide to the degree of selection on the unobservables. Researchers often informally argue for the exogeneity of an explanatory variable or an instrumental variable by examining the relationship between the instrumental variable and a set of observed characteristics, or by assessing whether point estimates are sensitive to the inclusion of additional control variables.\(^1\) We provide a formal theoretical analysis confirming the intuition that such evidence can be informative in some situations. More importantly, we provide ways to quantitatively assess the degree of selection bias or omitted variables bias and in some situations provide ways to estimate bounds. Our use of observables is not a strategy that can substitute for point identification, but rather a generalization that results in bounds. If there is a lot of “selection on the observables” then

the bounds can be very wide, but in the ideal case in which there is very little selection on observables and the explanatory power of the observables is high, the bounds will be tight. To fix ideas, let the $Y_i$ be a continuous outcome of interest determined by:

\begin{equation}
Y_i = \alpha T_i + X_i' \Gamma_X + W_i^* \Gamma^* + \xi_i
\end{equation}

where $T_i$ is a treatment variable.\(^2\) The parameter of interest is $\alpha$, the causal effect of $T_i$ on $Y_i$. $X_i$ is a vector of observed variables with coefficient vector $\Gamma_X$. $X_i$ contains variables that are always observed, and $W_i^*$ is a vector of additional characteristics that are relevant for determining the outcome which may or may not be observable to the econometrician. The final term, $\xi_i$, represents idiosyncratic shocks that are unrelated to the other components in the model. We use the notation $W_i' \Gamma$ to refer to the observed components of $W_i^* \Gamma^*$ and $W_i^u' \Gamma^u$ to refer to its unobserved components. We can rewrite the model as:

\begin{equation}
Y_i = \alpha T_i + X_i' \Gamma_X + W_i' \Gamma + (W_i^u' \Gamma^u + \xi_i)
\end{equation}

with the term in parentheses capturing all the unobservable components of the outcome.

The key idea in this paper is to model the relationship between an instrument $Z_i$ (for which a special case is $T_i$) and $W_i^u$. Our operational definition of "selection on unobservables is like selection on observables" is that the partial correlations of $W_i' \Gamma$ and $W_i^u' \Gamma^u$ with the instrument $Z_i$ are the same. The motivation for this involves thinking about the breakdown of exactly which characteristics are in $W_i$ versus $W_i^u$ as being determined by random chance. In addition, we view both $W_i$ and $W_i^u$ as having a large number of elements, none of which dominates in determining $Y_i$.\(^3\) Dominant characteristics, like gender or schooling in a wage regression, are assumed always measured and in $X_i$. Finally, although the principal source of endogeneity bias here is that $Z_i$ is correlated with $W_i^u$, an additional source of bias stems from the correlation between $W_i$ and $W_i^u$. In the context of a model for the determination of $W$, the correlations between the elements of $W_i$ are informative about the nature of the correlation between $W_i$ and $W_i^u$.

\(^2\)We will also discuss a binary dependent variable model in which the outcome is $1(Y_i > 0)$.

\(^3\)We will utilize approximations that take the number of regressors in $W^*$ (and $W$) to be large.
To illustrate the nature of the restrictions we use, consider the linear projection of $T_i$ onto $X_i$, $W_i^\prime \Gamma$ and $W_i^w \Gamma^u$:

\[(1.3) \text{Proj}(Z_i | X_i, W_i^\prime \Gamma, W_i^w \Gamma^u) = \phi_0 + X_i^\prime \phi_X + \phi W_i^\prime \Gamma + \phi_u W_i^w \Gamma^u.\]

Our formalization of the idea that, after controlling for $X_i$, “selection on the unobservables is the same as selection on the remaining observables” is that:

**Condition 1.**

\[\phi_u = \phi.\]

One may contrast Condition 1 with the implication of the usual IV orthogonality conditions:

**Condition 2.**

\[\phi_u = 0.\]

Roughly speaking, Condition 1 says that conditional on $X_i$, the part of $Y_i$ that is related to the observables and the part related to the unobservables have the same relationship with $Z_i$. Condition 2 says that the part of $Y_i$ related to the unobservables has no relationship with $Z_i$.

A projection like that in equation (1.3) will only be directly useful when an approximation for $W_i^w \Gamma^u$ is available. When $\xi_i$ is nonzero, the composite term $(W_i^w \Gamma^u + \xi_i)$ is all that can be approximated. The analog of equation (1.3) is

\[(1.4) \text{Proj}(Z_i | X_i, W_i^\prime \Gamma, (W_i^w \Gamma^u + \xi_i)) = \phi_0 + X_i^\prime \phi_X + \phi W_i^\prime \Gamma + \phi_u (W_i^w \Gamma^u + \xi_i).\]

With some abuse of notation, we continue to use $\phi_u$ as the last coefficient. Equal partial correlations of $Z_i$ with $W_i^\prime \Gamma$ and $W_i^w \Gamma^u$ in this projection will imply an inequality of $\phi$ and $\phi_u$ due to attenuation bias in the latter coefficient. This results in an intermediate condition 3 between the extremes of Conditions 1 and 2, defined as:

**Condition 3.**

\[0 \leq \phi_u \leq \phi \text{ if } \phi \geq 0\]

\[0 \geq \phi_u \geq \phi \text{ if } \phi < 0.\]
We propose two alternative estimators use Condition 3. They differ in how they model the relationship between \( W_i \) and \( W_i^u \). We call the first estimator OU, which refers to using properties of observed ("O") covariates to infer the properties of unobserved ("U") covariates. OU amounts to estimating equation (1.2) using moment conditions that \( X \) and \( W_i \) are orthogonal to \( W_i^u \) and the restrictions in Condition 3. This estimates bounds on \( \alpha \). It requires a high level assumption that implies, roughly speaking, that conditional on \( X_i \), the coefficient of the regression of \( Z_i \) on \( (Y_i - \alpha T_i) \) has the same sign and is at least as large in absolute value as the coefficient of the regression of the part of \( Z_i \) that is orthogonal to \( W_i \) on the part of \( Y_i - \alpha T_i \) that is orthogonal to \( W_i \). The high level assumption is required because the estimator does not make direct use of how the observed and unobserved explanatory variables are interrelated to assess the consequences of omitted variables that affect both the treatment and the outcome. Essentially, it treats \( W_i \) as exogenous, in common with the vast IV literature that focusses on endogeneity of \( T_i \) but treats the "controls" as exogenous. Furthermore, it does not provide a way to account for the fact that randomness in which elements of \( W_i^* \) are observed influences the distribution of the estimator. This estimator has been applied in Altonji, Elder and Taber (2005a, 2005b; hereafter, AET) to study the effectiveness of Catholic schools, as well as in a large number of other studies.\(^4\) We complete the theoretical analysis of the estimator that is presented in preliminary form in AET (2002).

We also propose a second estimator that we believe is a more satisfactory approach because it relaxes the assumption that \( W_i \) is exogenous. In this second approach, we develop a method of moments procedure that uses the bounds on selection embodied in Condition 3 and also uses a factor structure to model the covariance between the observable and unobservable covariates. This structure allows us to infer properties of unobserved covariates based on the observed correlation structure of the observed covariates \( W_i \). We show that this estimator, which we name OU-Factor, consistently identifies a set that contains \( \alpha \). We also provide a general bootstrap procedure that may be used to construct confidence regions for the identified set, as well as a less computationally demanding bootstrap procedure that seems to works well in practice.

\(^4\)AET also propose a way to measure the amount of selection on the index of observables that determine the outcome and then calculate a ratio of how large selection on unobservables would need to be in order to attribute the entire OLS estimate of \( \alpha \) to selection bias. This approach, which has been applied in a number of subsequent papers, is closely related to the OU estimator.
Our paper is related in spirit to the rapidly growing emphasis in econometrics on partial identification and bound estimation. Some of these papers implicitly address omitted variables and selection bias. Rosenbaum and Rubin (1983) and Rosenbaum (1995) propose examining the sensitivity of \( \alpha \) to varying \( \phi_u \). As we’ve already noted, our paper has antecedents in the very large number of papers that examine the link between \( T_i \) (or an instrumental variable or an regression discontinuity indicator in an IV or regression discontinuity context) to the other covariates that influence \( Y_i \) and use the pattern as qualitative evidence about whether \( T_i \) is likely to be correlated with the omitted variables that influence \( Y_i \). Our contribution is the development of a formal model of how the observed variables relate to the unobserved variables and the translation of the informal intuition that the patterns in the observables are informative about the unobservables into bounds estimators.\(^5\)

The paper continues in Section 2, where we provide a formal model of which covariates are observed and which are unobserved. We provide an explicit set of assumptions under which Condition 1, Condition 2, and Condition 3 hold, and we elaborate on why Condition 3 is the most plausible of the three. In Section 3 we present the OU estimator. We also show that in general, Condition 1 is not sufficient to provide point identification of \( \alpha \). As a practical matter, this is not critical, because we focus on the use of Condition 3 to identify a range of admissible values for \( \alpha \). We then turn to the OU-Factor estimator based on specifying a factor structure for \( W_i^f \). In Section 4 we provide some Monte Carlo evidence on the performance of OU and OU-Factor. We offer brief conclusions in Section 5.

2 Selection Bias and the Link Between the Observed and Unobserved Determinants of the Instrument and Outcome

In this section, we begin with a formal discussion of how the observables \( W_i \) are chosen from the full set \( W_i^* \). This is the first step in developing a theoretical foundation for using the relationship between a potentially endogenous variable (or an instrument for that variable) and the observables to make inferences about the relationship between such a variable and

\(^5\)A large literature on survey non-response and to item nonresponse that leads to missing data on dependent variables or covariates for some observations, of which Kline and Santos (2010) is a recent example. We ignore item non-response and focus on missing variables.
the unobservables. In doing so, we provide a foundation for quantitatively assessing the importance of the bias from the unobservables. We then provide a set of conditions under which Condition 3 holds, which is central to OU and OU-factor.

2.1 How are Observables Chosen?

We do not know of a formal discussion of how variables are chosen for inclusion in data sets. Here we make a few general comments that apply to many social science data sets. First, most large scale data sets such as the National Longitudinal Survey of Youth 1979, the British Household Panel, the Panel Study of Income Dynamics, and the German Socioeconomic Panel are collected to address many questions. Data set content is a compromise among the interests of multiple research, policy making, and funding constituencies. Burden on the respondents, budget, and access to administrative data sources serve as constraints. Obviously, content is also shaped by what is known about the factors that really matter for particular outcomes and by variation in the feasibility of collecting useful information on particular topics. Major data sets with large samples and extensive questionnaires are designed to serve multiple purposes rather than to address one relatively specific question. As a result, explanatory variables that influence a large set of important outcomes (such as family income, race, education, gender, or geographical information) are more likely to be collected. Because of limits on the number of the factors that we know matter, that we know how to collect, and that we can afford to collect, many elements of $W^*_i$ are left out. This is reflected in the relatively low explanatory power of most social science models of individual behavior. Furthermore, in many applications, the treatment variable $T_i$ is correlated with many of the elements of $W^*_i$.

These considerations suggests several conclusions about data set content. First, at least some of a small set of exogenous variables that play a critical role in determining $Y_i$ and $T_i$ are likely to be available in data sets that were designed with a particular research topic in question. These variables are represented by $X_i$. In AET’s study of Catholic schools, Catholic religion is such a variable.

Second, Condition 2, which underlies single equation methods in econometrics, will rarely hold in practice even though the optimal survey design for estimation of $\alpha$ would be to assign the highest priority to variables that are important determinants of both $T_i$ and $Y_i$. (It would
also be to useful to collect potential instrumental variables that determine $T_i$ but not $Y_i$.) Condition 2 is based on the extreme assumption that surveys are sufficiently well designed to ensure that $\phi_u = 0$.

Third, it may be better to think of the variables that are available as somewhat random. Indeed, an alternative to Condition 2 which is, in a sense, the other extreme from it, is to assume that the constraints on data collection are sufficiently severe that it is better to think of the elements of $W_i$ as an approximately random subset of the elements of $W_i^*$, rather than being systematically chosen to eliminate bias. Putting aside $X_i$, a natural way to formalize the idea that “selection on the observables is the same as selection on the unobservables” is to treat observables and unobservables symmetrically by assuming that the observables are a random subset of a large number of underlying variables. Let the indicator $S_j$ denote whether covariate $j$ is observed in the data set. A symmetric treatment of observables and unobservables in our framework is to assume that $S_j$ is an iid binary random variable which is equal to one with probability $P_{S_j}$ for all covariates in $W_i^*$. (Later we will discuss the case in which $P_{S_j}$ differs across the elements of $W_i^*$.)

Finally, there are many reasons to include idiosyncratic shocks $\xi_i$ in the framework. In many problems outcomes are determined considerably after the treatment $T_i$, characteristics $X_i$, or instruments $Z_i$ are determined. Consider the case of the effect of deciding to attend Catholic high schools on 12th grade test scores studied by AET. All of the regressors used in AET are measured in eighth grade. High school outcomes will be influenced by shocks that occur during the four years of high school, many of which are unanticipated at the time of decision regarding whether to attend a Catholic school. Given this sequencing, these shocks influence high school outcomes but cannot affect the probability of starting a Catholic high school. In addition, $\xi_i$ will be needed to reflect random variability in a student’s performance which has nothing to do with the decision to attend Catholic high school. Similarly, in health applications, $\xi_i$ may reflect health shocks (such as an accident or exposure to a virus) that occur after the treatment choice $T_i$ has been made.

### 2.2 Implications of Random Selection of Observables

We are now ready to consider the implications of random covariate selection from $W_i^*$. We begin with the general case. We first derive the probability limit of $\phi_u/\phi$ as the number of
covariates in $W^*_i$ becomes large. We then consider several special cases.

We define outcomes as being determined by a sequence of models indexed by $K^*$, where $K^*$ is the number of elements of $W^*_i$. A natural part of the thought experiment in which $K^*$ varies across models is the idea that the importance of each individual factor declines with $K^*$. We take the dimension of $X_i$ as fixed.

Define $G^{K^*}$ as the information set consisting of the realizations of the $S_j$, coefficients $\Gamma_j$, and the joint distribution of $W_{ij}$ conditional on $j = 1, \ldots, K^*$. That is, $E(W_{ij} | G^{K^*})$ is the mean for a given $j$, where the expectation is only over $i$, but $E(W_{ij})$ is an unconditional expectation over both $i$ and $j$. It may be helpful to think of this data generation process as operating in two steps. First the “model” is drawn: for a given $K^*$, the joint distribution of $W_{ij}, T_i, Z_i, \xi_i$, and $S_j$ are drawn. $G^{K^*}$ represents this draw. In the second stage of the data generating process, individual data are constructed from these underlying distributions.

The two steps combine to generate $Y_i$ as is represented in Assumption 1.

**Assumption 1.**

$$Y_i = \alpha T_i + X'_i \Gamma X + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij} \Gamma_j + \xi_i$$

where $(W_{ij}, \Gamma_j)$ is unconditionally stationary (indexed by $j$), and $X_i$ includes an intercept.

We use (and slightly abuse) non-standard notation in Assumption 1. Rather than explicitly indexing parameters by $K^*$, we suppress a $K^*$ index on $(W_{ij}, \Gamma_j)$ and bring a $\frac{1}{\sqrt{K^*}}$ out in front of the sum. This scaling guarantees that no particular covariate will be any more important *ex ante* than the others. It embodies the idea that a large number of components determine most outcomes in the social sciences.\(^6\) Any variables that play an outsized role in $Y_i$ and $Z_i$ are assumed to always be in the set of special regressors $X_i$. Note that Assumption 1 involves unconditional stationarity. Conditional on $G^{K^*}$, the variance of the $W_{ij}$ and the contribution of the $W_{ij}$ to the variance of $Y_i$ will differ across $j$.

---

\(^6\)The “local to unity” literature in time series econometrics (e.g., Stock, 1994) and the “weak instruments” literatures (e.g., Staiger and Stock, 1997) are other examples in econometrics in which the asymptotic approximation is taken over a sequence of models, which in the case of those literatures, depend on sample size. However, in those literatures the purpose of the model sequence is provide a better guide to the asymptotic distribution of estimator, which is quite different from the present case.

\(^7\)It does restrict the model sequence to the class in which the ratio of the effects of $W_{ij}$ and $W_{ij'}$ for all $j, j' < K^*$ to be independent of $K^*$.
Throughout we will project all variables on $X_i$ and take residuals to remove $X_i$ from the regression. We will use “tildes” to denote the residuals from these projections, so we define

$$
\tilde{W}_{ij} \equiv W_{ij} - \text{Proj}(W_{ij} \mid X_i; G^{K^*})
$$

$$
\tilde{T}_i \equiv T_i - \text{Proj}(T_i \mid X_i; G^{K^*})
$$

$$
\tilde{Z}_i \equiv Z_i - \text{Proj}(Z_i \mid X_i; G^{K^*})
$$

$$
\tilde{Y}_i \equiv Y_i - \text{Proj}(Y_i \mid X_i; G^{K^*})
$$

where $\text{Proj}$ denotes a linear projection.\(^8\) Let $\sigma_{j,\ell}^{K^*} = E(\tilde{W}_{ij} \tilde{W}_{i\ell} \mid G^{K^*})$. To guarantee that $\text{var}(Y_i)$ is bounded as $K^*$ becomes large, we assume that

**Assumption 2.**

$$
0 < \lim_{K^* \to \infty} \frac{1}{K^*} \sum_{j=1}^{K^*} \sum_{\ell=1}^{K^*} E(\sigma_{j,\ell}^{K^*} \Gamma_j \Gamma_\ell) < \infty ; \quad \lim_{K^* \to \infty} \text{Var} \left( \frac{1}{K^*} \sum_{j=1}^{K^*} \sum_{\ell=1}^{K^*} \sigma_{j,\ell}^{K^*} \Gamma_j \Gamma_\ell \right) \to 0 .
$$

The next assumption guarantee that $\text{cov}(\tilde{Z}_i, \tilde{Y}_i)$ is well behaved as $K^*$ grows.

**Assumption 3.** For any $j = 1, ... , K^*$, define $\mu_j^{K^*}$ so that

$$
E \left( \tilde{Z}_i \tilde{W}_{ij} \mid G^{K^*} \right) = \frac{\mu_j^{K^*}}{\sqrt{K^*}} .
$$

We assume that

$$
E(\mu_j^{K^*} \Gamma_j) < \infty ; \quad \lim_{K^* \to \infty} \text{Var} \left( \frac{1}{K^*} \sum_{j=1}^{K^*} \mu_j^{K^*} \Gamma_j \right) \to 0 .
$$

In section A.1 of Appendix A we prove that Assumptions 2 and 3 are satisfied by an MA(q) model. In section A.7 we show that the factor model for $\tilde{Z}_i$ and $\tilde{W}_{ij}$, which is central to the OU-Factor estimator also satisfies these restrictions.

The key assumption provides defines the process under which observables are chosen. Consider the case discussed above in which variables are chosen at random:

---

\(^8\)Formally, the linear projection projection of a generic $Y_i$ on a generic $X_i$ is defined by $X_i^T \delta$ where $\delta$ satisfies $E[(Y_i - X_i^T \delta)X_i \mid G^{K^*}] = 0$. Hereafter, this projection is meant to be the population projection conditional on $G^{K^*}$, i.e., for a very large $N$, but with $K^*$ draw of $G^{K^*}$ and fixed.
Assumption 4. For \( j = 1, \ldots, K^* \), \( S_j \) is independent and identically distributed with \( 0 < \Pr (S_j = 1) \equiv P_s \leq 1 \). \( S_j \) is also independent of all other random variables in the model. If \( \text{var}(\xi) \equiv \sigma^2 = 0 \), then \( P_s < 1 \).

Finally we need an assumption on the remaining error term \( \xi_i \).

Assumption 5. \( \xi_i \) is mean zero and uncorrelated with \( \tilde{Z}_i \) and \( \tilde{W}_{ij} \).

The assumption that \( \xi_i \) is uncorrelated with \( \tilde{W}_{ij} \) is not very restrictive, since for a given value of \( K^* \) one can redefine \( \Gamma^* \) and \( \xi \) so that \( \xi \) is uncorrelated with \( W^* \).

First we consider the relationship between \( \phi \) and \( \phi_u \) in the general case with nonzero \( \text{var}(\xi_i) \) and then derive three key special cases.

Note that our asymptotic analysis is nonstandard in two respects. First, we are allowing the number of underlying explanatory variables, \( K^* \), to get large. Second, the random variable \( \tilde{W}_{ij} \) is different from the random variables \( \Gamma_j \) and \( S_j \) in the following way. For each \( j \) we draw one observation on \( \Gamma_j \) and \( S_j \) which is the same for every person in the population; however, each individual \( i \) draws her own \( \tilde{W}_{ij} \).

Theorem 1. Define \( \phi \) and \( \phi_u \) such that

\[
\text{Proj} \left( \tilde{Z}_i \mid \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} \Gamma_j, \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij} \Gamma_j + \xi_i; G^{K^*} \right) = \phi \left( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} \Gamma_j \right) + \phi_u \left( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij} \Gamma_j + \xi_i \right).
\]

Then under assumptions 1-3 and 4-5, if the probability limit of \( \phi \) is nonzero, then

\[
\frac{\phi_u}{\phi \overset{p}{\rightarrow} \frac{(1 - P_s)A}{(1 - P_s)A + \sigma^2}} \quad \text{as } K^* \rightarrow \infty.
\]

Assume one can write \( \xi_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_j B_{w_j} + \tilde{\xi}_i \) where \( \tilde{\xi}_i \) is mean independent of \( W^* \) and \( Z \). Replace \( \Gamma_j \) with \( \Gamma_j + B_{w_j} \) and replace \( \xi_i \) with \( \tilde{\xi}_i \) in Assumption 1. The key assumption would then be that \( \tilde{\xi}_i \) is uncorrelated with unobservables that determine \( Z_i \), which we define to be \( \psi_i \) below.
where

\[ A \equiv \lim_{K^* \to \infty} E \left( \frac{1}{K^*} \sum_{j=1}^{K^*} \sigma_{j,j}^2 \Gamma_j^2 \right). \]

If the probability limit of \( \phi \) is zero, then the probability limit of \( \phi_u \) is also zero.

(Proof in section A.2 of Appendix A)

Next we consider three separate cases which we present as corollaries. These are the formal analogies of Conditions 1, 2, and 3. We omit the proofs of these as they follow immediately from the proof of Theorem 1.

**Corollary 1.** When \( \sigma^2_\xi = 0 \),

\[ \text{plim}(\phi - \phi_u) = 0. \]

Corollary 1 states that the coefficients of the projection of \( \tilde{T}_i \) onto \( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} \Gamma_j \) and \( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij} \Gamma_j \) approach each other with probability one as \( K^* \) becomes large. The other extreme is the case in which all the important control variables that affect both \( \tilde{Z}_i \) and \( \tilde{Y}_i \) are included in the model, so the variation in the composite error term \( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij} \Gamma_j + \xi_i \) arises from \( \xi_i \) only:

**Corollary 2.** When \( P_s = 1 \),

\[ \text{plim}(\phi_u) = 0. \]

What about the case in which selection on observables is stronger than selection on unobservables but there is still some selection on unobservables? This corresponds to the case in which \( \text{var}(\xi) > 0 \) and \( P_s < 1 \). The next Corollary considers this case:

**Corollary 3.** When \( 0 < P_s < 1 \) and \( \sigma^2_\xi > 0 \),

either

\[ 0 < \text{plim}(\phi_u) < \text{plim}(\phi), \]

or

\[ \text{plim}(\phi) < \text{plim}(\phi_u) < 0, \]
or
\[ 0 = \text{plim}(\phi_u) = \text{plim}(\phi). \]

This Corollary plays a key role in the estimator below.

### 2.3 Systematic Variation in \( P_{S_j} \)

In this subsection we extend Theorem 1 to the case in which \( Pr(S_{ij} = 1) \) is positively related to the impact of including \( \tilde{W}_{ij} \) on the bias in IV estimation of \( \alpha \). Without loss of generality, assume the correlation between \( \tilde{Z}_i \) and \( \tilde{W}_i'\Gamma \) is positive (as one could multiply \( \tilde{Z}_i \) by -1 to change the sign). In general, the impact of including a particular \( \tilde{W}_{ij} \) is a complicated function of \( \Gamma_j, \mu_j, \Gamma_\ell \) and \( \mu_\ell \) of the variables that remain excluded, and the covariances among both the included and excluded variables. Thus, it is not straightforward to characterize the relative impact of the exclusion of particular variables on the bias. Nevertheless, Theorem A.1 in Appendix A (section A.3) provides a necessary and sufficient condition for \( 0 < \phi_\varepsilon < \phi \). The condition is complicated and since it is not central to developing the estimators below, in the text we focus on a special case that satisfies the condition. Assume that \( S_j \) is independent of \( \tilde{W}_{ij}\Gamma_j \):

**Assumption 6.** \( S_j \) is independent of \( \tilde{W}_{ij}\Gamma_j \).

This is neither an attractive assumption nor a necessary condition, but it implies that the variation in \( P_{S_j} \) will not affect the second moments of

\[
\left\{ \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij}\Gamma_j, \left( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij}\Gamma_j + \xi_i \right) \right\}
\]

as \( K^* \) gets large.\(^{10}\) Under this additional assumption, Theorem 2 implies that \( 0 < \phi_\varepsilon < \phi \) if

\[
E(\mu_j\Gamma_j \mid S_j = 1) > E(\mu_j\Gamma_j \mid S_j = 0) \geq 0.
\]

It is intuitive that exclusion of \( \tilde{W}_{ij} \) variables with a strong positive association with both \( Z_i \) and \( Y_i \) will lead to bigger bias, everything else equal.

\(^{10}\)If we impose (3.1), we might instead assume \( S_j \) is positively correlated with \( E(\beta_j\Gamma_j) \) but unrelated to the marginal distributions of \( \Gamma_j, \beta_j, \) and \( \text{cov}(W_j, W_\ell) \) for all \( j \) and \( \ell \).
A key, but perhaps subtle implication of this special case and of the more general result in Theorem 2 is that we get the inequality \( \phi_\varepsilon < \phi \) even when \( \sigma_\xi^2 = 0 \). Even in the original model with constant \( P_{S_j} = P_s \) for \( 1 \leq j \leq K^* \), there is systematic variation in the probability that a determinant of \( Y \) will be included in the sense that the variables that determine the index \( \xi_i \) cause the least bias (0) and have the lowest probability (0) of being included. This underlies Theorem 1.

3 Estimators of \( \alpha \)

We now discuss ways to estimate \( \alpha \). In Section 4.1 We set the stage by reviewing the OU estimator introduced in AET (2002, 2005). Then we present OU-Factor, beginning with the factor model of \( \tilde{W}_i^c \) that it requires.

But before turning to the estimators, we provide an explicit model for \( \tilde{Z}_i \) which we use for both estimators.

Assumption 7.

\[
\tilde{Z}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \tilde{W}_{ij} \beta_j + \psi_i,
\]

where (i) \( \psi_i \) is independent of all of the elements of \( \tilde{W}_i^c \). (ii) \( \beta_j \) is a stationary process with finite second moments. \( \beta_j \) may be correlated with \( \Gamma_j \).

It is convenient to rewrite the model for \( \tilde{Z}_i \) as

\[
(3.1) \quad \tilde{Z}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K} \tilde{W}_{ij} \beta_j + u_i
\]

where \( u_i = \frac{1}{\sqrt{K^*}} \sum_{j=K+1}^{K^*} \tilde{W}_{ij} \beta_j + \psi_i \), and all variables are residuals from linear projections onto the space of \( X_i \).

3.1 The OU Estimator

We repeat the outcome equation here for convenience

\[
\tilde{Y}_i = \alpha \tilde{T}_i + \tilde{W}_i' \Gamma + (\tilde{W}_i' \Gamma + \xi_i)
\]

\[
\equiv \alpha \tilde{T}_i + \tilde{W}_i' \Gamma + \varepsilon_i
\]
Instrumental variables estimation of $\alpha$ uses the standard moment conditions $E(\tilde{W}_i \varepsilon_i) = 0$ and the IV moment equation $E(\tilde{Z}_i \varepsilon_i) = 0$. The simplest form of the OU estimator replaces the moment equation $E(\tilde{Z}_i \varepsilon_i) = 0$ with Condition 3. In most applications of OU to date either $T_i = 1(Z_i > 0)$ or $T_i = Z_i$, so we focus on the case:

$$T_i = \tilde{Z}_i = \frac{1}{\sqrt{K}} \tilde{W}'_i \beta + u_i.$$  

A problem, however, is that mean independence of $\varepsilon_i$ and $\tilde{W}$ is not likely to hold. Mean independence is maintained in virtually all observational studies of selection problems because without it, $\alpha$ is not identified even if one has a valid exclusion restriction.\footnote{The exception is when the instrument is uncorrelated with $W_i$ (and $X_i$) as well as $\xi_i$, as when the instrument is randomly assigned in an experimental setting.} Our discussion of how the observables are determined makes clear that mean independence of $\varepsilon_i$ and $\tilde{W}$ is hard to justify in most settings-including ours. If the observables are correlated with one another, as in most applications, then the observed and unobserved determinants of $Y_i$ are also likely to be correlated. This will lead to an inconsistent estimator whether one uses $E(\tilde{Z}_i \varepsilon_i)$ as a moment condition or one uses Condition 3. Note that this does not reflect a problem with Theorem 1. Theorem 1 did not require $E(\tilde{W}_i \varepsilon_i) = 0$ because it involves the true value of $\Gamma$. Here we need an assumption analogous to $E(\tilde{W}_i \varepsilon_i) = 0$ in order to consistently estimate $\Gamma$ or provide an alternative condition in the absence of a consistent estimate.

To address the problem of endogenous $\tilde{W}_i$, OU relies on an additional high level assumption, while OU-factor addresses the problem head on by modeling the link between the observed and unobserved $\tilde{W}_j$. In the OU case, AET assume that $E(\varepsilon_i | \tilde{W}_i)$ is linear, and define $G$ and $e$ to be the slope vector and error term of the “reduced forms”:

$$E\left(\tilde{Y}_i - \alpha \tilde{T} | \tilde{W}\right) \equiv \tilde{W}'G$$ (3.2)

$$\tilde{Y} - E\left(\tilde{Y}_i - \alpha \tilde{T} | \tilde{W}\right) \equiv e.$$ (3.3)

Let $\phi_{W'G}$ and $\phi_e$ be the coefficients of the projection of $T$ on $W'G$ and $e$ in a regression model that includes $X$. Note that under the assumption that if $E(\tilde{W}_i \varepsilon_i) = 0$, then $G = \Gamma$ and under the assumptions of Theorem 1, $0 \leq \phi_e \leq \phi_{W'G}$ when $\phi_{W'G} > 0$. AET show that this is true under the following more general (though not necessarily easy to interpret) condition:
Assumption 8.

\[
\lim_{K^* \to \infty} \frac{\sum_{\ell=1}^{K^*} E \left( \tilde{W}_{ij} \tilde{W}_{ij-\ell} \right) E \left( \beta_j \Gamma_{j-\ell} \right)}{\sum_{\ell=1}^{K^*} E \left( \tilde{W}_{ij} \tilde{W}_{ij-\ell} \right) E \left( \Gamma_j \Gamma_{j-\ell} \right)} = \lim_{K^* \to \infty} \frac{\sum_{\ell=1}^{K^*} E \left( \tilde{\tilde{W}}_{ij} \tilde{\tilde{W}}_{ij-\ell} \right) E \left( \beta_j \Gamma_{j-\ell} \right)}{\sum_{\ell=1}^{K^*} E \left( \tilde{\tilde{W}}_{ij} \tilde{\tilde{W}}_{ij-\ell} \right) E \left( \Gamma_j \Gamma_{j-\ell} \right)},
\]

for the set of variables \( W_{ij} \) in \( j = 1, \ldots, K^* \),

where \( \tilde{\tilde{W}}_{ij} \) is the component of \( \tilde{W}_{ij} \) that is orthogonal to the observed variables \((X_i, W_i)\), for all elements of \( W^*_i \). Roughly speaking Assumption 8 says that the coefficient of the regression of \( \tilde{T}_i \) on \( \left( \tilde{Y}_i - \alpha \tilde{T}_i - \xi_i \right) \) is equal to the coefficient of the regression of the part of \( \tilde{T}_i \) that is orthogonal to \( \tilde{W}_i \) on the corresponding part of \( \left( \tilde{Y}_i - \alpha \tilde{T}_i - \xi_i \right) \). This condition holds under the standard assumption \( E(\varepsilon_i | \tilde{W}_i; G^{K^*}) = 0 \), in which case \( G \) and \( e_i \) equal \( \Gamma \) and \( \varepsilon_i \), respectively. However, \( E(\varepsilon_i | \tilde{W}_i; G^{K^*}) = 0 \) is not necessary for Assumption 8.\(^{12}\)

**Theorem 2.** Define \( \phi_{\tilde{W}'G} \) and \( \phi_e \) such that

\[
\begin{aligned}
\text{Proj} \left( \tilde{Z}_i \mid \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} G_j, \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1-S_j) \tilde{\tilde{W}}_{ij} \Gamma_j + \xi_i; G^{K^*} \right) \\
= \phi_{\tilde{W}'G} \left( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} \Gamma_j \right) + \phi_e \left( \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1-S_j) \tilde{\tilde{W}}_{ij} \Gamma_j + \xi_i \right).
\end{aligned}
\]

Then under assumptions 1-5 and 8, as \( K^* \) gets large, then

\[
\phi_e \xrightarrow{p} \frac{\sum_{\ell=-\infty}^{\infty} E \left( \tilde{W}_{ij} \tilde{W}_{ij-\ell} \right) E \left( \Gamma_j \Gamma_{j-\ell} \right)}{\sum_{\ell=-\infty}^{\infty} E \left( \tilde{W}_{ij} \tilde{W}_{ij-\ell} \right) E \left( \Gamma_j \Gamma_{j-\ell} \right) + \sigma^2_i}
\]

if the probability limit of \( \phi \) is nonzero. If the probability limit of \( \phi_{\tilde{W}'G} \) is zero then the probability limit of \( \phi_e \) is also zero.

(Proof in Section A.4 of Appendix A)

The upshot is that one can work with the system

\(^{12}\)For example, one can show that Assumption 8 will also hold if \( E \left( \beta_j \Gamma_{j-\ell} \right) \) is proportional to \( E \left( \Gamma_j \Gamma_{j-\ell} \right) \) regardless of the correlations among the \( W_j \).
\[ \tilde{Y}_i = \alpha \tilde{T}_i + \frac{1}{\sqrt{K^*}} \tilde{W}'_i G + e_i. \]
\[ \tilde{T}_i = \frac{1}{\sqrt{K^*}} \tilde{W}'_i \beta + u_i \]

and estimate the set of \( \alpha \) values that satisfy the above inequality restrictions. In practice, AET find that the lower bound is obtained when the equality of selection condition
\[
\frac{\text{cov}(u_i, e_i | G^K_i)}{\text{var}(e_i | G^K_i)} = \frac{\text{Cov}(\tilde{W}'_i \beta, \tilde{W}'_i G | G^K_i)}{\text{Var}(\tilde{W}'_i G | G^K_i)}
\]
is imposed and the upper bound corresponds to the case in which \( \tilde{T}_i \) is treated as exogenous, with \( \frac{\text{cov}(u_i, e_i | G^K_i)}{\text{var}(e_i | G^K_i)} = 0 \).

One can perform statistical inference accounting for variation over \( i \) conditional on which \( W_i \) are observed in the usual way, and we omit the details. However, there is no obvious way to account for random variation due to the draws of \( S_j \). This is another reason one might prefer OU-factor

### 3.2 The OU-Factor Estimator

#### 3.2.1 A Factor Model of \( \tilde{W}_{ij} \)

The biggest issue with the OU estimator is that it requires Assumption 8 which in general is hard to justify in a model in which the \( W_{ij} \) are chosen randomly from the set of \( W_i^* \).

Relaxing this assumption requires building a model of the relationship between the \( W_{ij} \) that we observe and the \( W_{ij} \) that we do not observe. We do this using a factor model of \( \tilde{W}_{ij} \), which is central to the estimator proposed below. The factor model is a convenient way to model the relationship among the covariates. We assume that

**Assumption 9.** \( \tilde{W}_{ij} \) has a factor structure

\[ \tilde{W}_{ij} = \frac{1}{\sqrt{K^*}} \tilde{F}'_i \Lambda_j + \nu_{ij}, \ j = 1, \ldots, K^*, \]

where \( \tilde{F}_i \) is an \( r \) dimensional mean zero vector of factors.

We treat \( r \) as finite, so while the dimension of \( \tilde{W}_{ij} \) grows, the number of factors remains constant. Recall that \( \tilde{W}_{ij} \) is the residual from the projection of \( W_{ij} \) upon \( X_i \). \( \tilde{F}_i \) is the
residual from the projection of the factors that determine \( W_{ij} \) on \( X_i \). We normalize the variance/covariance matrix of \( \tilde{F}_i \) be to the identity matrix. Define \( \sigma_j^2 \equiv E(v_{ij}^2 \mid \mathcal{G}^{K^*}) \), \( j = 1, \ldots, K^* \). It is important to contrast our work with other models using factor structures such as Cuhna, Heckman, and Schennach (2010). In much of this other work the factors themselves drive outcomes. Our model is quite different. We assume that outcomes are determined by \( \tilde{W}_{ij} \) itself. We use the factor structure only as a model of the covariance structure of \( W^*_i \).

Let the model for \( \tilde{T}_i \) be

**Assumption 10.**

\[
\tilde{T}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} \delta_j + \left[ \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij} \delta_j + \omega_i \right].
\]

For convenience repeat the equation for \( \tilde{Z}_i \) and \( \tilde{Y}_i \):

\[
(3.4) \quad \tilde{Z}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} \beta_j + \left[ \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij} \beta_j + \psi_i \right],
\]

\[
(3.5) \quad \tilde{Y}_i = \alpha \tilde{T}_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \tilde{W}_{ij} \Gamma_j + \left[ \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) \tilde{W}_{ij} \Gamma_j + \xi_i \right].
\]

The \( \omega_i \), \( \xi_i \) and \( \psi_i \) are assumed independent of all of the \( \tilde{W}_{ij} \). The instrument error term \( \psi_i \) is assumed to be correlated with the treatment error \( \omega_i \) but not the outcome error \( \xi_i \). The brackets in each of the above equations collect unobservable terms. Note that if all the elements of \( W^*_i \) were observed \( (S_j = 1 \text{ for all } j) \), our framework reduces to the standard instrumental variables setup.

The stochastic structure of the model is that \( \Lambda_j \), \( \Gamma_j \), \( \beta_j \) and \( \sigma_j^2 \) differ across \( j \), but are identical for all individuals in the population. We redefine \( \mathcal{G}^{K^*} \) to refer to aspects of the model of \( \tilde{W}_{ij}, \tilde{T}_i, \tilde{Y}_i, \) and \( \tilde{Z}_i \), that do not vary across individuals:

\[
\mathcal{G}^{K^*} = \{(\Gamma_j, \beta_j, \delta_j, \Lambda_j, \sigma_j^2, S_j) \text{ for } j = 1, \ldots, K^*\}.
\]

For estimation, we make the following additional assumptions.
Assumption 11. (i) \((\Gamma_j, \beta_j, \delta_j, \Lambda_j, \sigma_j)\) is i.i.d with fourth moments; (ii) the support of \(\sigma_j\) is \([\sigma, \tilde{\sigma}]\) with \(\sigma > 0\); (ii)the support of \(|\Lambda_j|\) is bounded above by \(\Lambda < \infty\) and \(E(\Lambda_j) \neq 0\); (iii) the error terms \((\omega_i, \psi_i, \xi_i)\) are mean zero with finite second moments and are independent of \(\tilde{W}_i^*\); (iv) \(\psi_i\) and \(\xi_i\) are uncorrelated, \(\omega_i\) and \(\psi_i\) are correlated.

Assumption 11 (iii) allows \(\tilde{T}_i\) to depend on the component \(\omega_i\) that is correlated with the instrument \(\tilde{Z}_i\) but uncorrelated with the observed and unobserved determinants of \(\tilde{Y}_i\). This is required for identification of \(\alpha\) in the standard IV case in which all elements of \(W_i^*\) are observed. In Appendix A (section A.5) we verify that the factor model of \(\tilde{W}_i\) in conjunction with the model assumptions 1, 7, and 10 satisfy Assumption 2 and 3 of Theorem 1.

3.2.2 The OU-factor Estimator of an Admissible Set for \(\alpha\)

In general the model is not point identified, so we provide an estimator of a set that contains the true values. The key subset of the parameter vector of our model is \(\theta = \{\alpha, P_s, \sigma^2_{\xi}\}\). The treatment effect \(\alpha\) is the main parameter of interest. \(P_s\) is the probability that \(S_j = 1\), \(\sigma^2_{\xi}\) is the variance of \(\xi_i\). The true value of \(\theta\) is \(\theta_0 = \{\alpha_0, P_{s0}, \sigma^2_{\xi0}\}\) which lies in the compact set \(\bar{\Theta}\). Our approach is to estimate a set \(\hat{\Theta}\) that asymptotically will contain the true value \(\theta_0\). The key restrictions on the parameter set are

\[
0 < P_{s0} \leq 1, \text{ and} \\
\sigma^2_{\xi0} \geq 0.
\]

The case in which \(P_{s0} = 1\) is the standard IV case represented by Condition 2, while \(\sigma^2_{\xi0} = 0\) is the “unobservables like observables” case represented by Condition 1. We construct an estimate of the set of values of \(\alpha\) by estimating the set of \(\theta\) that satisfy all of the conditions and then projecting onto the \(\alpha\) dimension. (In Section 3.4 we discuss construction of confidence intervals.) While the upper and lower bound of the estimated set does not have to correspond to the cases in which \(P_{s0} = 1\) and \(\sigma^2_{\xi0} = 0\), in practice we find that it does.

It will be helpful to make use of matrix notation. We assume that the variables are ordered so that \(j = 1, \ldots, K\) corresponds to the \(K\) observed covariates in \(W^*\). Unless indicated otherwise,

- For a generic variable \(B_i, i = 1, \ldots, N\), \(B\) will represent the \(N \times 1\) vector.
• For a generic variable \( B_j, j = 1, ..., K^* \), \( B \) will represent the \( K \times 1 \) vector of observable characteristics and \( B^* \) will represent the full \( K^* \times 1 \) vector.

• For a generic variable \( B_{ij}, i = 1, .., N, j = 1, ..., K^* \), \( B \) will represent the \( N \times K \) matrix of observable characteristics, \( B^* \) the full \( N \times K^* \) matrix of covariates, and \( B_i \) represents the \( K \times 1 \) vector of \( B_{ij} \) for a given \( i \).

• We also employ the convention of using capital letters for matrices so, for example, the matrix version of \( v_{ij} \) will be written as \( V \).

Given the large amount of notation we concentrate on the 1 factor case \((r = 1)\), so \( \hat{\Phi}_i \) and \( \Lambda_j \) are scalars. We fully expect that the results generalize to the multiple factor case. We now present the estimator, which has two stages.

**Stage 1**

In the first stage we estimate the terms of the factor loading. Since \( \Lambda \) is normalized by \( K^* \), but \( K^* \) is not observed, in this stage we define \( \lambda = \sqrt{P_{S0}} \Lambda \). We estimate \( \lambda \) as

\[
\hat{\lambda}_j \equiv \frac{\frac{1}{K} \sum_{\ell \neq j} 1_N \sum_{i=1}^N \hat{W}_{ij} \hat{W}_{i\ell}}{\sqrt{\frac{1}{K} \sum_{\ell_1} \sum_{\ell_2 \neq \ell_1} 1_N \sum_{i=1}^N \hat{W}_{i\ell_1} \hat{W}_{i\ell_2}}}.
\]

We then estimate \( \sigma_j^2 \) using

\[
\hat{\sigma}_j^2 \equiv \frac{1}{N} \sum_{i=1}^N \hat{W}_{ij}^2 - \frac{\hat{\lambda}_j^2}{K}.
\]

**Stage 2**

We estimate the rest of the parameters in a second stage. If we knew \( \alpha_0 \) and \( P_{s0} \) we could estimate \( \Gamma \) by taking advantage of the \( K \) moment conditions corresponding to the observable elements of \( W \),

\[
\sqrt{K^*} E \left[ \hat{W}_{ij} \left( \hat{Y}_i - \alpha_0 \hat{T}_i \right) \right] | G^{K^*} = \sqrt{K^*} E \left[ \left( \frac{1}{\sqrt{K^*}} \sum_{\ell=1}^{K^*} \hat{F}_i \Lambda_{ij} \Gamma_{\ell} + \frac{1}{\sqrt{K^*}} \sum_{\ell=1}^{K^*} v_{ij} \Gamma_{\ell} \right) + \xi_i | G^{K^*} \right]
\]

\[
= \Lambda_j \left( \frac{1}{K^*} \sum_{\ell=1}^{K^*} \Lambda_{ij} \Gamma_{\ell} \right) + \sigma_{ij}^2 \Gamma_j
\]

\[
\overset{p}{\rightarrow} \Lambda_j E(\Lambda_{ij} \Gamma_{\ell}) + \sigma_{ij}^2 \Gamma_j.
\]
We construct our estimator \( \hat{\Gamma} (\theta) \) with the sample analog of the above expression,
\[
\left[ \sqrt{K^*} \frac{1}{N} \tilde{W}' \left( \tilde{Y} - \alpha_0 \tilde{T} \right) \right] = \left[ \frac{1}{K} \frac{1}{P_s} \hat{\lambda} \hat{l} \hat{\Gamma} (\theta) + \Sigma \hat{\Gamma} (\theta) \right],
\]
where \( \Sigma \) is the diagonal matrix composed of the \( \sigma_j^2 \) terms. Thus, for the parameter \( \theta \) we can construct the estimator
\[
(3.9) \quad \hat{\gamma} (\theta) \equiv \left[ \frac{1}{P_s K} \hat{\lambda} \hat{l} + \hat{\Sigma} \right]^{-1} \frac{1}{N} \tilde{W}' \left( \tilde{Y} - \alpha \tilde{T} \right),
\]
where we define \( \hat{\Sigma} \) to be the diagonal matrix composed of the \( \hat{\sigma}_j^2 \), which is estimated in the first stage and we think of \( \hat{\gamma} (\theta) \) as an estimator of \( \hat{\Gamma} (\theta) / \sqrt{K^*} \).

To estimating the model we need to come up with estimating equations. We have not assumed anything about moments higher than order 2 and first moments are all zero once we have residualized the model. Thus the identifying information comes from second order moments. We have three fundamental parameters to estimate \( \theta = \{\alpha, P_s, \sigma_2^2\} \), but it turns out that we only have two different moments to identify them. In Appendix A (section A.6) we show that with \( K^* \) fixed the parameters are not identified as we are left with two moments.\(^{13}\)

The two equations that we use are
\[
q_{N,K^*}^1 (\theta) = \frac{1}{N} \sum_{i=1}^N Z_i \left( \tilde{Y}_i - \alpha \tilde{T}_i \right) - \frac{1}{K} \frac{\hat{\gamma} (\theta)' \hat{\lambda} \hat{l} \hat{\lambda}}{P_s} + \frac{\hat{\beta}' \hat{\Sigma} \hat{\gamma} (\theta)}{P_s}
\]
\[
q_{N,K^*}^2 (\theta) = \frac{1}{N} \sum_{i=1}^N \left( \tilde{Y}_i - \alpha \tilde{T}_i \right)^2 - \frac{1}{K} \left( \frac{\hat{\gamma} (\theta)' \hat{\lambda}}{P_s} \right)^2 - \frac{\hat{\gamma} (\theta)' \hat{\Sigma} \hat{\gamma} (\theta)}{P_s} - \sigma_2^2
\]
The first is analogous to the standard IV moment condition while the second is analogous to the total sum of squares. Under the true value of the parameters we show that these two equations converges in probability to zero as \( N \) and \( K^* \) get large.

That is, \( q_{N,K^*}^2 (\theta) \) is the difference between the total sum of squares of \( \left( \tilde{Y}_i - \alpha \tilde{T}_i \right) \) in the data for the hypothesized value of \( \alpha \) and the sum of squares implied by the model estimate.

We define the estimator \( \hat{\Theta} \) as the set of values of \( \theta \) that give sufficiently low values of the criterion function
\[
Q_{N,K^*} (\theta) = q_{N,K^*} (\theta)' \Omega q_{N,K^*} (\theta),
\]
\(^{13}\)We do not see any way that allowing \( K^* \) to grow will help this problem
conditional on $\widehat{\lambda}$ and $\widehat{\Sigma}$ from Stage 1 and subject to (3.6) and (3.7), where

$$q_{N,K^*}(\theta) = \begin{bmatrix} q_{1,N,K^*}(\theta) & q_{2,N,K^*}(\theta) \end{bmatrix}'$$

and $\Omega$ is some predetermined symmetric positive definite weighting matrix.

### 3.3 Consistency and Asymptotic Normality of the Estimator

In this section we prove consistency. Specifically we take joint limits as both jointly get large.

We maintain the assumptions of the factor model $W$ and Assumption (11). In addition we add Assumptions 12 and 14 below.

**Assumption 12.** $\bar{\Theta}$ is compact with the support of $P_s$ bounded below by $p_s^\ell > 0$.

**Assumption 13.** $K^*/N \to 0$ as $N$ grows

This last assumption is necessary for consistency of this particular estimator. It is likely the case that one could adapt the estimator in a way that will kill the $K^*/N$ bias terms and allow $K^*$ and $N$ to grow at the same rate. However, this assumption does not seem unreasonable given the sample sizes that social scientists normally use and thus seems to be the more interesting starting point.

**Assumption 14.** The dimension of $\tilde{F}_i$ is 1

**Theorem 3.** Under Assumptions 1, 4, 7, and 9-14, $Q_{N,K^*}(\theta)$ converges uniformly to $Q_0(\theta)$ and $Q_0(\theta_0) = 0$.

(Proof in Appendix A, Section A.9)

As described in Chernozhukov, Hong, and Tamer (2007), an issue with set identified models is precisely how to define the estimator and consistency of the estimator. As an example one could take $c$ as fixed and define the estimator as

$$\widehat{\Theta} = \{ \theta \in \bar{\Theta} : Q_{N,K^*}(\theta) < c \}.$$
Clearly from Theorem 3, $\hat{\Theta}$ is consistent in the sense that

$$Pr\left(\theta_0 \in \hat{\Theta}\right) \to 1.$$  

We can asymptotically get smaller sets $\hat{\Theta}$ by letting $c$ decline with the sample size using the rates of convergence below.

One can form a set estimator for $\alpha_0$ just by taking the projection of $\hat{\Theta}$ onto $\alpha$. That is, we can define this set as

$$\hat{A} \equiv \{ \alpha : \text{there exists some value of } (P_s, \sigma^2_\xi) \text{such that } \{\alpha, P_s, \sigma^2_\xi\} \in \hat{\Theta} \}$$

To get the asymptotic distribution we need to strengthen the assumptions about convergence. There are bias terms of order $K^*/N$ in the estimator above. In order to guarantee that these go away when we multiply by $\sqrt{K^*}$ we need the assumption that

**Assumption 15.** $\frac{K^*^3}{N^2} \to 0$ as $N$ grows

The asymptotic distribution turns out to be quite simple in this case

**Theorem 4.** Under Assumptions 1,4,7, 9-12,14, and 15

$$\sqrt{K^*q_{N,K^*}^1} (\theta_0) = -E\left(\sigma^2_j \Gamma_j \beta_j\right) \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (S_j - P_{S0}) + o_p(1)$$

$$\sqrt{K^*q_{N,K^*}^2} (\theta_0) = -E\left(\sigma^2_j \Gamma_j^2\right) \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (S_j - P_{S0}) + o_p(1)$$

(Proof in Appendix A, Section A.12)

**3.4 Constructing Confidence Intervals**

In this section we discuss confidence interval construction. We start with the ideal procedure one would use given unlimited computing resources. We then discuss a more practical approach, which we use in the Monte Carlos below.
3.4.1 A General Procedure

Before discussing inference it is useful to step back and consider our basic approach. In terms of identification we have three parameters \((\alpha_0, P^0_S, \sigma^0_\xi)\) but only 2 equations: the population and limit of the sequence of models for \((q^1_N, q^2_N)\).\(^{14}\) However, we also have limits on the parameter space. In particular \(0 < P_S \leq 1\) and \(\sigma^0_\xi \geq 0\). While we cannot get a point estimator for \((\alpha_0, P^0_S, \sigma^0_\xi)\), we construct the set estimator \(\hat{\Theta}\) for this three dimensional parameter. Our set estimate for \(\alpha_0\) is just the set of \(\alpha\) that lie within this identified set.

We can construct a confidence region in the analogous manner. That is, we could first construct a confidence set for \((\alpha_0, P^0_S, \sigma^0_\xi)\) and then let our confidence set for \(\alpha_0\) be the values of \(\alpha\) that lie within this set. The most natural way to construct the larger confidence set would be to “invert a test statistic.” That is, we would first construct a test statistic \(T(\theta)\) which has a known distribution under the null hypothesis: \(\theta = \theta_0\). For each potential \(\theta\), we would construct an acceptance region of the test. When \(T(\theta)\) lies within this acceptance region, \(\theta\) would belong to this confidence set, otherwise it would not. Given the confidence set for the full parameter space, we take the confidence set to be the set of \(\alpha\) that lie within this set. More formally let \(T_{N,K^*}(\theta)\) be the estimated value of the test statistic and let \(T^c(\theta)\) the critical value such that we reject \(\theta = \theta_0\) when the test statistic is larger than the critical value. The confidence set is defined as

\[
\hat{C}_{N,K^*} = \{ \theta \in \Theta \mid T_{N,K^*}(\theta) \leq T^c(\theta) \},
\]

and our estimated confidence region for \(\alpha\) can be written as

\[
\hat{C}_\alpha = \left\{ \alpha \in \mathbb{R} \mid \exists (P_S, \sigma_\xi) : (\alpha, P_S, \sigma_\xi) \in \hat{C}_{N,K^*} \right\}.
\]

There are many test statistics one could use and many ways to calculate the critical value. We consider the following algorithm based on the bootstrap. Consider testing the null hypothesis \(\theta = \theta_0\). The most natural test statistic is the normalized criteria function, so that

\[
T_{N,K^*}(\theta_0) = q_{N,K^*}(\theta_0)' \Omega q_{N,K^*}(\theta_0).
\]

\(^{14}\)In the definition of the estimator, we have not explicitly defined \(\Lambda, \Gamma, \beta,\) or \(\Sigma\) as parameters but express the estimates of these objects as functions of the data and \(\theta\). Because the dimension of these objects grows with \(K^*\), it is easier to focus on the elements of \(\theta\) when considering consistency and inference.
We then use a bootstrap procedure to calculate the critical value. As we need to define the full data generation process we assume that

**Assumption 16.**

\[
X_i = \Lambda_X F_i + v_{ix}
\]

\[
W_{ij} = \frac{1}{\sqrt{K^*}} F_i \Lambda_j + v_{ij}, \; j = 1, ..., K^*
\]

where \( v_{ix} \) is uncorrelated with everything else in the model as is \( F_i \).

This implies our factor model above where

\[
\tilde{F}_i = F_i - \text{proj}(F_i|X_i; G^{K^*})
\]

We calculate the bootstrap distribution of \( \sqrt{K^*} q_{N,K^*}(\theta_0) \) using the following procedure:

1. Estimate parameters to be used in generating data for the bootstrap. This involves using the data generation process for \( X_i \) as well. Specifically, from the empirical distribution of \( (X_i, W_i) \), for a given \( \theta \)
   
   (a) Estimate \( \Lambda, \Lambda_X, \Sigma \), and the data generating processes for \( F_i, v_{ix}, \) and \( v_{ij} \).
   
   (b) Estimate

   \[
   \hat{\gamma}(\theta_0) = \left[ \frac{1}{P_S K} \hat{\lambda}' \hat{\lambda} + \hat{\Sigma} \right]^{-1} \frac{1}{N} \tilde{W}' \left( \tilde{Y} - \alpha_0 \tilde{T} \right)
   \]

   \[
   \hat{b}(\theta_0) \equiv \left[ \frac{1}{P_S K} \hat{\lambda}' \hat{\lambda} + \hat{\Sigma} \right]^{-1} \frac{1}{N} \tilde{W}' \tilde{Z}
   \]

   (c) For the hypothesized value of \( P_S \), estimate the distribution of \( (\xi_i, \psi_i) \).

2. Generate \( (N_B, K_B^*) \) bootstrap samples as follows for each sample.

   (a) Draw \( K^* \) observable covariates from the actual set of covariates (with replacement) with appropriate \( \left( \hat{\Gamma}_j, \hat{\beta}_j, \hat{\lambda}_j, \hat{\sigma}_j \right) \). Here \( \hat{\lambda}_j = \frac{1}{\sqrt{P_S}} \hat{\lambda}_j \) and \( \hat{\Gamma}_j \) and \( \hat{\beta}_j \) are the jth components of \( \hat{\gamma}(\theta_0) \) and \( \hat{b}(\theta_0) \) multiplied by \( \sqrt{K^*} \).

   (b) For each \( j \) in that sample, draw a binary variable \( S^b_j \) where \( Pr(S^b_j = 1) = P_S \).
(c) Now for $i = 1, N$ generate all of the $(X_i, W_i^*)$ using the DGP for $F_i$, $v_{ij}$ and $v_{ix}$.

(d) Using the DGP for $\psi_i$ and $\xi_i$ generate $Z_i$ and $(Y_i - \alpha_0 T_i)$ (Note that we do not need to generate data on $Y_i$ and $T_i$ separately because only $\left(\tilde{Y}_i - \alpha_0 \tilde{T}_i\right)$ enters the moment conditions that define the test statistic.)

(e) Given generated bootstrap data construct $q_{b,N,K^*}$ using the empirical approach defined in section 3.2 above. (This involves the intermediate steps of estimating $\Sigma, \lambda$ and $\Gamma$ as well.)

3. From the bootstrap sample we can estimate the distribution of the test statistic and calculate the critical value given the size of the test.

For this critical value to be consistent, the bootstrap distribution of $T_{N,K^*}(\theta_0)$ must provide a consistent estimate of the actual distribution of $T_{N,K^*}(\theta_0)$.

**Theorem 5.** Under Assumptions 1,4,7,10-12, and 14-16, the bootstrap distribution of $\sqrt{K^*}q_{b,N,K^*}$ defined in the procedure above converges in distribution to the asymptotic distribution of $\sqrt{K^*}q_{N,K^*}(\theta_0)$.

(Proof in Appendix A, Section A.16)

The computational burden of computing $T_c(\theta)$ for the desired confidence level is likely to be very large. However, the moments that determine the criterion function of the model are continuous functions of $\theta$. Consequently, $T_c(\theta)$ should be a smooth function of $\theta$. We propose computing a modest number of draws of $Q_{N,K^*}(\theta)$ for each of the grid points of $\theta$ chosen and then approximating $T_c(\theta)$ by fitting a quantile regression model to the draws for the various values of $\theta$. One can increase the number of grid points, number of draws, and the flexibility of the quantile regression model as needed to ensure that the approximation is accurate for the confidence level chosen. The restrictions $0 < P_S \leq 1$ and $0 \leq \sigma^2_\xi < var(\tilde{Y}_i)$ as well as the fact that the sign of $\phi_u$ is known in some applications reduces the number of points that must be entertained.
3.4.2 A Simplified Bootstrap Procedure

Given the computational complexity of the above procedure, we also propose a less demanding alternative. An additional motivation for the alternative procedure stems from the fact that one often has a strong prior about the sign of the selection bias. We can obtain tighter bounds by imposing this prior (formally defined as “monotone selection” in Manski and Pepper, 2000). While our estimation interval can potentially be much more complicated, in simulations we consistently find a compact region with one end of the region occurring at the instrumental variable estimate \( P_S = 1 \) and the other occurring at the “observables like unobservables” assumption \( \sigma_\xi = 0 \). Without loss of generality we will assume positive selection bias so that the upper bound occurs under the constraint \( P_S = 1 \). We will also assume that the minimum value occurs at \( \sigma_\xi \). We propose a parametric bootstrap procedure to construct one-sided confidence interval estimators for the lower and upper bounds of this set, denoted \( \alpha_{\min} \) and \( \alpha_{\max} \), respectively. For concreteness, suppose one chooses a confidence level of \((1 - \phi)\). We construct these intervals such that the estimator \( \hat{\alpha}_{\phi, \max} \) has the nominal probability \( \phi \) of being below \( \alpha_{\min} \). The estimator \( \hat{\alpha}_{\phi, \max} \) has the nominal probability \( \phi \) of exceeding \( \alpha_{\max} \).

3.4.3 Construction of \( \hat{\alpha}_{\phi, \max} \)

The procedure for estimating \( \hat{\alpha}_{\phi, \max} \) involves the following steps.

1. Estimate the model parameters under the assumption that \( \sigma_\xi = 0 \) by solving the system of equations
\[
0 = q_1^1(\hat{\alpha}_{\min}, \hat{P}_S, 0) = q_2^2(\hat{\alpha}_{\min}, \hat{P}_S, 0)
\]
for \( \hat{\alpha} \), and \( \hat{P}_S \). In doing this we also obtain estimates of \( \Lambda, \Lambda_X, \Sigma, \Sigma_X \) and \( \gamma \) for \( X \) and the observable \( W_j \).

2. Next estimate some additional parameters that will be used for generating the bootstrap sample.

(a) Obtain estimates of the distributions for \( F_i, v_{ij}, \) and \( v_{xi} \) given the estimates of \([\hat{\Sigma}, \hat{\Lambda}_j]\). This can be done in a number of different ways. One could specify a parametric distribution and estimate the distribution parameters. Alternatively, one
could do this completely nonparametrically. A third possibility is to take advantage of the fact that our estimator involves up to second moments of the variables, so only up to 4th moments of the distributions of these variables matter for the sampling distribution of \( \hat{\alpha}_{\min} \). Instead of specifying parametric distributions, one could use a method of moments procedure to estimate up to the fourth moments from sample estimates of \( E(\tilde{W}_{ij}^r\tilde{W}_{ij}^s) \) and \( \tilde{\sigma}_v, \tilde{\Lambda}_j, j = 1, ..., K \) for various values of \( r \) and \( s \). One could then pick convenient parametric distributions for \( F_i \) and \( v_{ij}, j = 1, ..., K \) and choose parameters of the distributions to match the relevant moments.\(^{15} \) Call the estimates of the additional parameters of the \( F_i \) distribution \( \hat{B}_F \) and the additional parameters of the \( v_{ij} \) distribution \( \hat{B}_{v_i} \).\(^{16} \) A similar procedure can be used to estimate additional parameters \( \hat{B}_{Vx} \) of the distribution of the vector \( v_{xi} \).

(b) Next we need to estimate the distribution of \( (\xi_i, \psi_i, \omega_i) \). We can use the same three approaches as in the previous case. To use the third we need estimates of fourth moments. To obtain them, one can use the fourth moments of \( \tilde{Y}_i - \tilde{\alpha}\tilde{T}_i, \tilde{Z}_i \) and \( \tilde{T}_i \). Consider

\[
E(\xi_i^4) = E(\tilde{Y}_i - \tilde{\alpha}\tilde{T}_i)^4 - E\left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \tilde{W}_{ij}\Gamma_j\right)^4 - E\left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \tilde{W}_{ij}\Gamma_j\right)^2\tilde{\sigma}_\xi^2.
\]

We have the estimate of \( \hat{\alpha}_{\min} \), so \( E(\tilde{Y}_i - \tilde{\alpha}\tilde{T}_i)^4 \) can be replaced with the corresponding sample moment. We also have estimates of \( E\left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \tilde{W}_{ij}\Gamma_j\right)^2 \) and \( \tilde{\sigma}_\xi^2 \). One can use a similar procedure to estimate \( E(\psi_i^4) \). The relevant moment condition is

\[
E(\psi_i^4) = E(\tilde{Z}_i)^4 - E\left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \tilde{W}_{ij}\beta_j\right)^4 - E\left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \tilde{W}_{ij}\beta_j\right)^2\tilde{\sigma}_\psi^2.
\]

\(^{15} \)Sticking with the one factor case and taking \( W_{ij} \) to be mean zero, using independence of \( \theta_i \) and the \( v_{ij} \), and using the fact that \( \text{var}(\theta_i) = 1 \), the moments are \( E(W_{ij}^4) = \Lambda_i^4E(\theta_i^4) + E(v_{ij}^4) + 4\Lambda_i^2\sigma^2_{ij} \) and \( E(W^2_{ij}W_{ij}^2) = \Lambda_i^2\Lambda_i^2E(\theta_i^2) + \sigma^2_{ij}\sigma^2_{ij} \) for all \( j, j' \neq j \) pairs. The idea generalizes to the multiple factor case.

\(^{16} \)An alternative is to use the \( K \) observed \( W_j \), impose the estimates \( \hat{\Lambda}_j \) and the estimates of \( \tilde{\sigma}_{v_{ij}} \), choose parametric distributions for \( \theta_i, v_{i1}, ..., v_{iK} \), and fit the parameters of those distributions. The chosen distributions should not impose constraints on the second and fourth moments. In principle, one could work with nonparametric distributions with the variance constrained to match the \( \sigma^2_{v_{ij}} \). A nonparametric approach is unattractive from a computational point of view, and given that our estimators only involve first and second moments, it does offer any clear advantages.
Note that this requires an estimate of $\hat{\beta}$ and $\sigma^2_\psi$, but estimating these is analogous to estimating $\hat{\gamma}$ and $\sigma^2_\xi$ where the dependent variable is now $\tilde{Z}_i$ rather than $\tilde{Y}_i - \alpha \tilde{T}_i$. Estimation of $\delta, \sigma^2_\omega$ and $E(\omega^4_i)$ is analogous. We would then pick convenient parametric distributions for this joint distribution, and estimate parameters $B_{\xi,\psi,\omega}$. The joint distribution should not constrain the second and fourth moments unless one wishes to impose additional a priori information (such as normality) on it. We leave implicit the fact that $\hat{B}_{\xi,\psi,\omega}$ depends on $\hat{\alpha}_{\min}$.

3. Construct the Bootstrap sample. This involves a few different steps.

(a) Using the estimates $[\hat{\beta}_j, \hat{\Gamma}_j, \hat{\sigma}_v, \hat{\Lambda}_j, \hat{B}_j, j = 1, \ldots, K$, and the estimates $\hat{P}_S$, draw $K^*$ values of $[\hat{\beta}_j, \hat{\Gamma}_j, \hat{\sigma}_v, \hat{\Lambda}_j, \hat{B}_j]$ by sampling with replacement from the $K$ estimated values. Let the first $K$ correspond to the “observed” $W$’s for purposes of the bootstrap replication.

(b) Using $(\hat{\sigma}_v, \hat{\Lambda}_j, \hat{B}_j)$ and $\hat{B}_F$, generate $(F_i)^{(b)}$, $(v_{ij})^{(b)}$ and then $W_{ij}^{(b)}$, $i = 1, \ldots, N$, $j = 1, \ldots, K^*$ where $(b)$ denotes the bth bootstrap replication, $(b) = 1, \ldots, N_{boot}$.

(c) Using the $K^*$ values of $\hat{\beta}_j$, the associated $K^*$ vectors $W_{ij}^{(b)}$, $\hat{\alpha}_{\min}$, and the draws of $\psi_i^{(b)}$, use $\hat{B}_{\xi,\psi,\omega}$ to generate $N$ values of $(Z_i^{(b)}, T_i^{(b)}, Y_i^{(b)})$.

4. For each bootstrap sample compute $\hat{\alpha}_{\min}^{(b)}$ by solving

$$0 = q^1_{N(b)}(\hat{\alpha}_{\min}^{(b)}, \hat{P}_S, 0) = q^2_{N(b)}(\hat{\alpha}_{\min}^{(b)}, \hat{P}_S, 0)$$

on the bootstrap samples.

5. Calculate the $\varphi^{th}$ quantile of the bootstrap sample of $\hat{\alpha}_{\min}$ and subtract the difference between that and our point estimate from our point estimate of $\hat{\alpha}_{\min}$ to obtain the lower bound of our confidence set.

3.4.4 Construction of $\hat{\alpha}_{\varphi,\max}$

To obtain $\hat{\alpha}_{\varphi,\max}$, we assume that the largest value of $\hat{\alpha}$ that satisfies the restrictions of the model is obtained when one imposes the assumption that $\hat{P}_S = 1$. This ignores the possibility that unobserved $\tilde{W}_{ij}$ that induce positive correlation between $\tilde{T}_i$ and $\tilde{Y}_i$. If one sets $\hat{P}_S$ to 1
in the matrix \( \frac{1}{P_{SK}} \hat{\lambda} \hat{\lambda} + \hat{\Sigma} \) and replaces the matrix with \( \tilde{W}'\tilde{W} \) in equation 3.9) for \( \Gamma(\hat{\theta}) \), then the solution for \( \hat{\alpha} \) is IV. Under the null, all of the \( W_j \) are observed. Thus we do not need to impose a model of how the \( W_j \) are related to each other to account for the effects of missing \( W_j \). One can construct the one sided confidence interval estimate using the appropriate robust standard error estimator given assumptions about serial correlation and heteroskedasticity in \( \xi_i \). Alternatively, one can use a conventional bootstrap procedure.

While the simplicity of the above approach is attractive, it has an important shortcoming. We have not been able to prove that OLS is the upper bound when \( P_S \) is less than 1 \( Cov(W,\varepsilon) \neq 0 \). This is because bias in \( \hat{\Gamma} \) may lead to a partially offsetting bias in \( \hat{\alpha} \).

4 Monte Carlo Evidence

(This section is somewhat out of date since our estimation procedure has changed somewhat since we did this analysis)

In this section we present Monte Carlo evidence on the performance of the lower bound estimator \( \hat{\alpha}_{\min} \) for the \( OU - Factor \) and \( \hat{\alpha}_{\max} \), which we estimate based on \( \hat{\alpha}_{OLS} \) because in our context \( \hat{\alpha}_{\max} \) turns out to be essentially the same as the OLS estimator.\(^{17}\) We also present evidence on the performance of the lower bound estimator for \( OU \), which we refer to \( \hat{\alpha}_{OU} \) in the tables. Finally, we demonstrate the feasibility of the general procedure for estimating confidence regions by computing estimates in 2 cases. We also provide monte carlo evidence confirming that the hypothesis tests of \( \hat{\theta} = \theta \) have approximately the correct size under the null \( \theta = \theta_0 \).

We assume that there are not \( X \) variables in the model \( (\Gamma_X = 0) \) so the equations of the model of \( Y_i, T_i, \) and \( W_{ij} \):

\(^{17}\)The OLS estimator is essentially the same as the estimate of \( \alpha \) based on our moment equations with \( P_S \) set to 1. The two differ because we use the moments implied by the estimated factor structure rather than the actual variance covariance matrix of \( W \) in the moment condition for \( \hat{\Gamma} \). In the designs we consider we found that the maximum value of \( \hat{\alpha} \) consistent with \( \sigma_\xi^2 > 0 \) occurred at \( P_S = 1 \), although we have not proved that this has to be the case for any model with a factor structure.
\[ Y_i = \alpha_0 T_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij} \Gamma_j + \xi_i \]

\[ = \alpha_0 T_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j W_{ij} \Gamma_j + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) W_{ij} \Gamma_j + \xi_i \]

\[ W_{ij} = \frac{1}{\sqrt{K^*}} F_i \Lambda_j + v_{ij} \]

\[ T_i = Z_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}^K \beta_j + \psi_i \]

We focus on the case in which \( F_i \) is a scalar \((r = 1)\). We vary assumptions about \( P_S \), the fraction of the \( W_{ij} \) variables that are included in the model.

### 4.1 W parameters

The distributions of the variables that determine \( W_{ij} \) are

\[ F_i \sim N(0, 1) \]

\[ v_{ij} \sim N(0, \sigma^2_{v_{ij}}); \quad \sigma_{v_{ij}} \sim U(1.0, 2.0) \]

\[ \Lambda_j = \bar{\Lambda} + \tilde{\Lambda}_j \]

\[ \tilde{\Lambda}_j \sim U(-\tilde{\Lambda}_{\text{max}}, \tilde{\Lambda}_{\text{max}}) \]

For this specification,

\[ E[Cov(W_j, W_{j'})|j \neq j'] = \frac{1}{K^*} E(\Lambda_j \Lambda_{j'}) = \frac{1}{K^*} \bar{\Lambda}^2 \] and

\[ E[Var(W_j)] = \frac{1}{K^*} \bar{\Lambda}^2 + \frac{1}{3K^*} [\bar{\Lambda}_{\text{max}}]^2 + E(\sigma_{v_{ij}}^2), \]

where the expectations are defined over \( j \) and \( j' \). We report \( \frac{E[Cov(W_j, W_{j'})]}{E[Var(W_j)]} \) in the tables below.

### 4.2 Parameters of the \( Y_j \) and \( T_j \) Equations

\( \Gamma_j \) and \( \beta_j \) have expected values \( \mu_{\Gamma} \) and \( \mu_{\beta} \), respectively, and depend on a common component \( \varepsilon_j \) and the components \( \varepsilon_{\Gamma_j} \) and \( \varepsilon_{\beta_j} \) that are specific to \( \Gamma_j \) and \( \beta_j \). They are determined by
\[ \Gamma_j = \mu_\Gamma + \frac{g_\varepsilon}{[g_\varepsilon^2 + (1 - g_\varepsilon)^2]^{5/2}} \varepsilon_j + \frac{(1 - g_\varepsilon)}{[g_\varepsilon^2 + (1 - g_\varepsilon)^2]^{5/2}} \Gamma_j \]

\[ \beta_j = \mu_\beta + \frac{b_\varepsilon}{[b_\varepsilon^2 + (1 - b_\varepsilon)^2]^{5/2}} \varepsilon_j + \frac{(1 - b_\varepsilon)}{[b_\varepsilon^2 + (1 - b_\varepsilon)^2]^{5/2}} \beta_j, \]

where \( \varepsilon_j, \varepsilon_{\Gamma_j}, \) and \( \varepsilon_{\beta_j} \) are uniform random variables with mean 0 and variance 1. They are mutually independent and independent across \( j \).

The parameters \( g_\varepsilon \) and \( b_\varepsilon \) determine relative weights on \( \varepsilon_j \) and the idiosyncratic terms \( \varepsilon_{\Gamma_j}, \varepsilon_{\beta_j} \), thereby determining the covariance between \( \Gamma_j \) and \( \beta_j \). The weights are normalized so that \( \text{var}(\Gamma_j) = \text{var}(\beta_j) = 1 \) regardless of the choice of \( g_\varepsilon \) and \( b_\varepsilon \). \( g_\varepsilon^2 \) and \( b_\varepsilon^2 \) are the shares of the variances accounted for by the common component \( \varepsilon_j \), respectively. For the above design,

\[ E(\Gamma_j \cdot \beta_{j'}) = \mu_\Gamma \mu_\beta + \frac{g_\varepsilon \cdot b_\varepsilon}{[[g_\varepsilon^2 + (1 - g_\varepsilon)^2]^{5/2} \cdot [b_\varepsilon^2 + (1 - b_\varepsilon)^2]^{5/2}], \ j = j' \]

\[ = \mu_\Gamma \mu_\beta, \ j \neq j' \]

\[ \text{cov}(\Gamma_j, \beta_{j'}) = \text{corr}(\Gamma_j, \beta_{j'}) = \frac{g_\varepsilon \cdot b_\varepsilon}{[[g_\varepsilon^2 + (1 - g_\varepsilon)^2]^{5/2} \cdot [b_\varepsilon^2 + (1 - b_\varepsilon)^2]^{5/2}], \ j = j' \]

\[ = 0, \ j \neq j'. \]

\[ E(\Gamma_j \cdot \Gamma_{j'}) = \mu_\Gamma^2 + 1, \ j = j' \]

\[ = \mu_\Gamma^2, \ j \neq j' \]

\[ E(\beta_j \cdot \beta_{j'}) = \mu_\beta^2 + 1, \ j = j' \]

\[ = \mu_\beta^2, \ j \neq j' \]

Below we consider the effects of varying \( g_\varepsilon \) and \( b_\varepsilon \), and we also consider a case in which \( \beta_j = 0 \) for all \( j \).

### 4.3 Additional Parameter Values

We also examine the sensitivity of the estimates to the importance of \( \psi \) and \( \xi \), the idiosyncratic components of \( T \) and \( Y \), respectively. To do this, we vary \( \sigma_\xi^2 \) so as to vary the expected
fraction of the variance of the unobservable component of \( Y \) that is due to \( \xi \). That is, we choose \( \sigma^2_\xi \) to manipulate

\[
R^2_\xi \equiv E \left[ \frac{\sigma^2_\xi}{K^* \text{Var} \left( \sum_{j=K_0+1}^{K^*} W_j \Gamma_j | \Gamma \right) + \theta \sigma^2_\xi} \right],
\]

where the expectation is defined over the joint distribution of \( \Gamma, \beta, \) and \( W \). Similarly, we set \( \sigma^2_\psi \) to control

\[
R^2_\psi \equiv E \left[ \frac{\sigma^2_\psi}{K^* \text{Var} \left( \sum_{j=1}^{K^*} W_j \beta_j | \beta \right) + \sigma^2_\psi} \right].
\]

We report \( R^2_\psi \) and \( R^2_\xi \) in the tables below. Note that for a given value of \( R^2_\xi \), the value of \( \sigma^2_\xi \) will depend on the choice of \( P_S \), but \( \phi \) and \( \phi_u \) will not. We view this as an attractive parameterization because we are primarily concerned with ensuring that \( \phi \) and \( \phi_u \) do not depend on \( P_S \). The expected values of \( \phi \) and \( \phi_u \) at the true \( \alpha \) are complicated functions of the parameters of the data generation process, so we simply compute the average values in each design as well as the average estimate of \( \hat{\phi} \) at \( \hat{\alpha}_{\text{min}} \). Note that the bias in OLS declines with \( P_S \) because \( \psi \) assumes an increasing important role as the source of variance in \( T_i \) that is orthogonal to the observed \( W_j \). However, the variance of \( \psi \) also rises when the covariance among the \( W_j \) is increased and when we change \( \mu_\beta \).

For all experiments, we set \( N = 2000 \) and report results based on 1000 Monte Carlo replications. The bootstrap estimates of the .10 one-sided confidence interval estimate is based on 1000 bootstrap replications for each Monte Carlo replication. We set \( K^* \) to 100 and \( \alpha_0 \) to 1.0 in all the experiments reported, and we set \( R^2_\psi \) to 0.5 in all experiments except Table 1, where it is set to 1. We vary \( P_S, R^2_\xi, \bar{\Lambda}, \bar{\Lambda}_{\text{max}}, \mu_B, \mu_\Gamma, g_\varepsilon, \) and \( b_\varepsilon \) across experiments. Specifically, we set \( P_S \) of 0.2, 0.4, and 0.8 and we set \( R^2_\xi \) to 0, 0.2, and 0.4. We vary \( \mu_B, \mu_\Gamma, \)

\[18\] If we fix \( \text{Var}(\xi_i) \) at a nonzero value, the ratio \( \phi_\varepsilon/\phi \) approaches 0 (the case in which OLS is unbiased) as \( P_S \) approaches 1. In assessing how variation in \( P_S \) matters, we wish to hold constant the degree to which selection on observables is similar to selection on unobservables. For each Monte Carlo experiment we set \( \sigma^2_\psi \) and \( \sigma^2_\xi \) to the fixed values

\[
\sigma^2_\xi = E \left[ \frac{R^2_\xi}{1 - R^2_\xi} \frac{1}{K^*} \text{Var} \left( \sum_{j=K_0+1}^{K^*} W_j \Gamma_j | \Gamma \right) \right],
\]

\[
\sigma^2_\psi = E \left[ \frac{R^2_\psi}{1 - R^2_\psi} \frac{1}{K^*} \text{Var} \left( \sum_{j=1}^{K^*} W_j \beta_j | \beta \right) \right],
\]

given the values of the other parameters of the experiment.
$g_\varepsilon$, and $b_\varepsilon$ such that $E(\beta_j \Gamma_j)$ takes on several different values. Finally, we vary $\bar{\Lambda}$ and $\tilde{\Lambda}_{\text{max}}$. In one set of case, we set $\bar{\Lambda} = 0$, which means that $E[\text{Corr}(W_{ij}, W_{ij'})] = 0$ if $j \neq j'$. In the other set of cases, $E[\text{Corr}(W_{ij}, W_{ij'})] = 0.2$ if $j \neq j'$.

### 4.4 Monte Carlo Results

We first consider a baseline case in which $T_i$ is randomly assigned. Table 1 reports results for a design in which $\beta_j = 0$ for all $j$ (i.e., $\mu_\beta = 0$, $\text{var}(\varepsilon_{\beta j}) = 0$, and $b_\varepsilon = 0$), which means that $T$ does not depend on the $W_j$. For these designs, $\hat{\alpha}_{\text{OLS}}$ is unbiased because $E(\phi) = E(\phi_u) = 0$.

We report the median as our measure of central tendency, and we also report the 10th and 90th percentile values as measures of dispersion. The median values of $\phi$, $\phi_u$, and $\tilde{\phi}$ across replications are shown in the top three rows of the table.

The estimates of $\hat{\alpha}_{\text{OLS}}$ are tightly distributed around 1.0 in all three cases. The dispersion declines with $P_S$, reflecting a smaller variance of the unobserved components of $Y$ as $P_S$ increases. The values of $\hat{\alpha}_{\text{OU}}$ and of $\hat{\alpha}_{\text{min}}$ are also tightly distributed around 1.0, although they are estimated less precisely than the OLS coefficients. When $P_S = 0.2$, the 90th-10th differential of $\hat{\alpha}_{\text{min}}$ is roughly double that of the 90th-10th differential for $\hat{\alpha}_{\text{OLS}}$, but when $P_S = 0.8$, the three estimators have similar dispersion.

We turn next to designs in which OLS estimates of $\alpha_0$ are biased. In Table 2a, we set $\mu_\beta = \mu_T = 0.3$, which leads to bias in $\hat{\alpha}_{\text{OLS}}$ in the specifications we consider. In the first three columns we chose $b_\varepsilon$ and $g_\varepsilon$ so that $E(\Gamma_j \beta_j) = 0.3$. The median of $\hat{\alpha}_{\text{OLS}}$ is 1.256 when $P_S = 0.2$ and 1.101 when $P_S = 0.8$. The decline in bias as $P_S$ increases reflects the fact that the fraction of the variance in $T_i$ that is uncorrelated with the excluded $W_j$ rises with $P_S$. $\hat{\alpha}_{\text{min}}$ is essentially unbiased in all three cases, with the dispersion declining with $P_S$. In the last three columns we increase $b_\varepsilon$ and $g_\varepsilon$ so that $E(\Gamma_j \beta_j) = 0.6$ (i.e., $\text{Corr}(\Gamma_j, \beta_j) = .51$). For each value of $P_S$, the bias in OLS increases relative to the cases in which $E(\Gamma_j \beta_j) = 0.3$. Interestingly, the $\hat{\alpha}_{\text{OU}}$ and $\hat{\alpha}_{\text{min}}$ estimators are less noisy compared to the $E(\Gamma_j \beta_j) = 0.3$ case. When $E(\Gamma_j \beta_j) = 0.6$ and $P_S = 0.8$, as shown in column 6, $\hat{\alpha}_{\text{OU}}$ and $\hat{\alpha}_{\text{min}}$ have no more sampling error than the OLS estimator.

Table 2b repeats the calculations found in Table 2a but introduces a factor structure such that $E[\text{Corr}(W_{ij}, W_{ij'})] = 0.2$ if $j \neq j'$. We impose this correlation by setting $\bar{\Lambda}$ to 3.4. In order to keep $E[\text{Var}(W_{ij})]$ constant relative to the $\bar{\Lambda} = 0$ case, we reduce $\tilde{\Lambda}_{\text{max}}$ from 6.2 to
2.0. The bias in OLS tends to be lower for this design, primarily because the regressors that are included do a better job of controlling for the omitted \(W_j\) when the correlation among the \(W_j\) is higher. Intuitively, as \(E[\text{Corr}(W_{ij}, W_{ij'})] \to 1\), it does not matter which regressors are actually observed and which are not. The increase in the correlation across \(W_j\) is also associated with an improvement in the performance of \(\hat{\alpha}_{\text{min}}\) relative to \(\hat{\alpha}_{OU}\). In particular, \(\hat{\alpha}_{OU}\) is downward biased in all of the designs apart from the one shown in the final column. This is likely due to the fact that the \(\hat{\alpha}_{OU}\) estimator is based on the assumption that the restriction \(\phi = \phi_u\) based on the true \(\Gamma_j\) carries over to the coefficient vector \(\Gamma^p\) of the projection of \(Y_i - \alpha_i T\) on the observables \(W_i\). However, the positive correlation between the observed and unobserved covariates results in positive omitted variables bias (on average) in the observed \(\hat{\Gamma}_j\), because the unobserved covariates are positively correlated with \(Y\). Since the observed covariates are also positively correlated with \(T\) in these designs, the positive bias on the estimates of \(\Gamma_j\) leads the projection of \(T\) on \(W_i \Gamma^p\) to overstate the amount of selection bias, inducing a negative bias in the \(\hat{\alpha}_{OU}\) estimates. This negative bias also affects the OLS estimator, partially counteracting the positive bias caused by the positive correlation of \(T\) with the unobserved elements of \(W\). As a result, the positive bias in the OLS estimates is smaller in Table 2b than in Table 2a.

As is evident from the table, \(\hat{\alpha}_{\text{min}}\) performs very well in the presence of a factor structure. It has a median value very close to 1 and a sampling error that is similar to OLS. Presumably, the superior performance of \(\hat{\alpha}_{\text{min}}\) relative to \(\hat{\alpha}_{OU}\) for the parameter values in Table 2b is due to the fact that explicitly accounting for the factor structure eliminates the positive bias on the estimates of \(\Gamma_j\), which in turn eliminates the negative bias in the estimate of \(\alpha_0\). However, the difference in performance between \(\hat{\alpha}_{\text{min}}\) and \(\hat{\alpha}_{OU}\) is only large in a few designs, such as those given by the first two columns in the table.

In Table 3, we relax the assumption that the observables are a random set of all the unobservables by setting \(R^2_\xi = 0.2\). In the left panel, \(\bar{A} = 0\) and \(\bar{\Lambda}_{\text{max}} = 6.2\), as in Table 2a. Not surprisingly, allowing for a positive variance of \(\xi\) has no effect on the median of OLS. However, the lower bound estimators \(\hat{\alpha}_{OU}\) and \(\hat{\alpha}_{\text{min}}\) are now both biased downward because the assumption that \(\phi = \phi_u\) no longer holds. This is easy to see in the first column, in which the median of \(\phi_u\) across replications is 0.353, roughly 80 percent of the median of \(\phi\) (0.438). In other words, selection on unobservables is now only 80 percent as large
as selection on observables. When \( E(\Gamma_j \beta_j) = 0.3 \) and the factor structure is such that \( E[\text{Corr}(W_{ij}, W_{ij'})] = 0 \), the medians of \( \hat{\alpha}_{OU} \) vary from 0.784 to 0.975 depending on \( P_S \), and the corresponding medians of \( \hat{\alpha}_{\text{min}} \) vary from 0.878 to 0.979. However, the sampling variance of the \( \hat{\alpha}_{OU} \) and \( \hat{\alpha}_{\text{min}} \) estimators is fairly wide when \( P_S \) is small. When we increase \( b_\varepsilon \) and \( g_\varepsilon \) so that \( E(\Gamma_j \beta_j) = 0.6 \), the positive bias in OLS increases, as was the case in Table 2a, while there is no systematic change for the other estimators. The sampling variances of \( \hat{\alpha}_{OU} \) and \( \hat{\alpha}_{\text{min}} \) are wider in this case than in the analogous cases in Table 2a (in which the assumption \( \phi = \phi_u \) holds). We do not fully understand this pattern, but in spite of it, the lower bound estimators usefully complement OLS.

The right panel of Table 3 sets \( \bar{\Lambda} \) and \( \tilde{\Lambda}_{\max} \) so that \( E[\text{Corr}(W_{ij}, W_{ij'})] = 0.2 \). The median values of \( \hat{\alpha}_{\text{min}} \) do not change very much relative to the case of independent \( W_j \), but the sampling distribution narrows substantially. This likely reflects the fact that when the \( W_j \) are correlated, it is easier to “fill in” for the effects of missing covariates using the OU-Factor moment conditions, so that it matters less which elements of \( W^* \) are actually observed.

Table 4 is analogous to Table 3, except now \( R_\xi^2 = 0.4 \), thereby lowering \( \phi_u \) relative to \( \phi \). The median of OLS is essentially unchanged relative to the cases in which \( R_\xi^2 \) is 0 or 0.2, which is not surprising. As one would expect, the medians of \( \hat{\alpha}_{OU} \) and \( \hat{\alpha}_{\text{min}} \) decline in all cases, with the largest declines occurring when \( P_S = 0.2 \). The medians of \( \hat{\alpha}_{\text{min}} \) range between 0.288 to 0.890 when \( E[\text{Corr}(W_{ij}, W_{ij'})] = 0 \). The sampling variability of the \( \hat{\alpha}_{OU} \) and \( \hat{\alpha}_{\text{min}} \) estimators also increases relative to Table 3. As expected, the sampling variance of \( \hat{\alpha}_{\text{min}} \) modestly improves when \( E[\text{Corr}(W_{ij}, W_{ij'})] \) increases from 0 to 0.2.

Table 5 summarizes an experiment in which \( \mu_\beta = 1 \), \( \mu_\Gamma = 5 \) and \( g_\varepsilon = b_\varepsilon = 0 \). For this specification \( E(\Gamma_j \beta_j) = 5 \), and \( \Gamma_j \) and \( \beta_j \) are uncorrelated. In the first three columns, \( E[\text{Corr}(W_{ij}, W_{ij'})] = 0 \) and \( R_\xi^2 = 0 \). OLS is badly upward-biased in these designs, with the median of \( \hat{\alpha}_{\text{OLS}} \) equaling 2.109 when \( P_S = 0.2 \), 1.929 when \( P_S = 0.4 \) and 1.419 when \( P_S = 0.8 \). The medians of \( \hat{\alpha}_{OU} \) and \( \hat{\alpha}_{\text{min}} \) range between 0.889 and 1.065, although they have a substantial sampling variance. In the middle three columns, \( E[\text{Corr}(W_{ij}, W_{ij'})] = 0.2 \). The bias in OLS declines but is still substantial when \( P_S = 0.2 \). Both \( \alpha_{OU} \) and \( \alpha_{\text{min}} \) perform well in these designs, as they are tightly distributed around the true value of \( \alpha \). In the last three columns we keep \( E[\text{Corr}(W_{ij}, W_{ij'})] = 0.2 \) and set \( R_\xi^2 = 0.2 \). The medians of \( \hat{\alpha}_{OU} \) and \( \hat{\alpha}_{\text{min}} \)
are roughly 0.26 when $P_S = 0.2$ and roughly 0.95 when $P_S = .8$, and the estimators have a relatively tight distribution. Overall, the designs in Table 5 highlight the fact that $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\text{min}}$ can perform very well in cases in which OLS is badly biased upward, particularly when $\phi = \phi_u$ holds. When $|\phi| > |\phi_u|$, so that selection on observables is stronger than selection on unobservables, the lower bound estimators yield values below the true $\alpha_0$, as expected, but the resulting bounds are often useful.

Finally, in Table 6 we explore the performance of the simplified bootstrap procedure for six designs described above. All results in the table are based on 1000 Monte Carlo replications, each of which includes 1000 bootstrap replications. The two columns in panel A correspond to columns 2 and 4 of Table 2a, in which $\bar{A} = 0$, so that $E[\text{Corr}(W_{ij}, W_{ij'})] = 0$, and $R^2_\xi = 0$. In the first column, in which $E(\Gamma_j \beta_j) = 0.3$, the empirical size, given by $\Pr(\hat{\alpha}_{0.10, \text{min}} < \alpha)$, is 0.087, based on a nominal size of 0.10. When $E(\Gamma_j \beta_j) = 0.6$, the empirical size equals 0.090, so that in both cases the confidence region given by $(\hat{\alpha}_{0.10, \text{min}}, \hat{\alpha}_{0.10, \text{max}})$ excluded $\alpha_0$ in slightly less than 10 percent of cases. The table also reports the median of the estimated standard error of $\hat{\alpha}_{\text{min}}$ across Monte Carlo replications, where the standard error in each replication is calculated across all 1000 bootstrap replications. In both cases, this estimated median standard error is larger than the standard deviation (across Monte Carlo replications) of $\alpha_{\text{min}}$. The fact that the bootstrapped distribution of $\hat{\alpha}_{\text{min}}$ is slightly more disperse than the analogous distribution across Monte Carlo replications is likely the cause of the under-rejection described above, i.e., that the empirical sizes of the tests are slightly smaller than the nominal size.

In panel B, $E[\text{Corr}(W_{ij}, W_{ij'})] = 0.2$, and designs in the two columns correspond to columns 2 and 4 of Table 2b. Again, coverage rates are close to the nominal size of 0.10, and median standard error estimate is in the ballpark of the standard deviation across replications of $\hat{\alpha}_{\text{min}}$.

Finally, in panel C, $R^2_\xi = 0.2$. In these cases, the estimated sampling variances of $\hat{\alpha}_{\text{min}}$ are slightly lower than the standard deviations across replications. While one might expect that this pattern would lead to over-rejection, i.e., empirical sizes greater than 0.10, the opposite case holds: in the first column, the empirical size is 0.038, and in the second it is only 0.001. This under-rejection occurs because the $\phi = \phi_u$ condition does not hold, so that the $\hat{\alpha}_{\text{min}}$ estimator is a conservative one – the lower bound given by $\hat{\alpha}_{\text{min}}$ will systematically
lie below $\alpha_0$, which is a restatement of the fact that the estimates of $\hat{\alpha}_{\text{min}}$ in Table 3 were biased downward. As a result, the confidence region given by $(\hat{\alpha}_{0.10,\text{min}}, \hat{\alpha}_{0.10,\text{max}})$ will include $\alpha_0$ in more than 90 percent of cases.

On the whole, the Monte Carlo results may be summarized as follows. First, the medians of $\hat{\alpha}_{\text{min}}$ and $\hat{\alpha}_{\text{OU}}$ are close to 1 when the assumption of equality of selection on observed and unobserved variables is correct ($R_\xi^2 = 0$). There are some differences in performance depending upon the specifics of the experiment, particularly the strength of the factor structure, but overall the two perform similarly. The sampling variances are narrower when the stronger is the factor structure, i.e., when $E[Corr(W_{ij}, W_{ij}')] = 0.2$. Second, both $\hat{\alpha}_{\text{min}}$ and $\hat{\alpha}_{\text{OU}}$ typically lie below the value of $\alpha_0$ when $\phi > \phi_u$. This is to be expected, because both estimators are based on the assumption that $\phi = \phi_u$ and are to be interpreted as lower bound estimators if $\phi > \phi_u > 0$ (in the case $\phi > 0$). Third, the gap between the lower bound estimators and $\alpha_0$ declines with $P_S$, which is also to be expected. Fourth, the $\hat{\alpha}_{\text{min}}$ and $\hat{\alpha}_{\text{OU}}$ estimators are usually less precise than is $\alpha_{\text{OLS}}$. The loss of precision depends on the design and is negligible in the case in which $T$ is randomly assigned (as in Table 1). For some designs, such as some of the cases with a strong factor structure in Table 2b, the sampling variance of $\hat{\alpha}_{\text{min}}$ is actually smaller than that of $\hat{\alpha}_{\text{OLS}}$. Overall, the distribution of $\hat{\alpha}_{\text{min}}$ and $\hat{\alpha}_{\text{OU}}$ are sufficiently precise to provide useful information about $\alpha$ in all of the cases that we consider.

5 Conclusion

In many situations, exclusion restrictions, functional form restrictions, or parameter restrictions are not sufficiently well grounded in theory or sufficiently powerful to provide a reliable source of identification. What can one do?

As we noted in the introduction, it is standard procedure to look for patterns in the relationship between an explanatory variable or an instrumental variable and the observed variables in the model when considering exogeneity. We provide a theoretical foundation for thinking about the degree of selection on observed variables relative to unobserved variables, and we propose two estimators that make explicit use of the pattern of selection in the observables to bound the treatment effect. We contrast the standard IV or OLS assumption
that the researcher has chosen the control variables so that the instrument (or the treatment itself) are not related to the unobservables with the assumption that the control variables are randomly chosen from the full set variables that influence the outcome, and argue that the truth is likely to lie somewhere in between.

Our estimators build on Theorem 1, which concerns the coefficients of the projection of an outcome on the regression indices of the observables and the unobservables. A number of assumptions are required, but roughly speaking, the theorem says that when the number of observed and unobserved variables that influence the outcome are large, the coefficient on the index of unobservables will lie between 0 and the coefficient on the index of observables. Both $OU$ and the $OU – Factor$ estimators identify bounds by imposing the inequality restriction on the econometric model for the outcome. However, in the likely case that the observed and unobserved variables are related, the coefficients on the control variables will to suffer from omitted variables bias, invalidating the restriction and the case for bounds. The $OU$ estimator combines Theorem 1 with a high level assumption about the link among the observed and unobserved variables. The $OU – Factor$ estimator adds the assumption that the observed and unobserved explanatory variables have a factor structure, which provides additional moment restrictions that permit one to account for the effects of omitted variables. We show that the estimator identifies a set that asymptotically contains the true value of the treatment parameter. We derive the asymptotic distribution of the $OU – Factor$ estimator and present a parametric bootstrap approach to statistical inference. Our Monte Carlo simulations are generally encouraging, particularly for $OU – Factor$.

There is a very long research agenda. More Monte Carlo evidence is needed in the context of real world applications and data sets. Thus far we have not applied the $OU – Factor$ estimator, and we have not performed Monte Carlo studies for designs with multiple factors. The $OU$ estimator has the advantage of simplicity and has already been used in a number of applications. However, a way to account for randomness in which explanatory variables are included in $W$ when constructing confidence intervals is needed. Ultimately, we believe that incorporating a formal model of the relationships among the observed and unobserved variables in $W^c$ is the more promising long-run research path. The linear factor model used in developing the $OU – Factor$ estimator is a natural way to do this, but it is also restrictive. Other models of the joint distribution of the covariates should be explored. We only touch
upon the case of heterogeneous treatment effects and so far we have only considered models in which the index that determines the outcome is an additively separable function.

More generally, we think of $OU$ and $OU - Factor$ as a start for an investigation into a broader class of estimators based on the idea that if one has some prior information about how the observed variables were arrived at, then the joint distribution of the outcome, the treatment variable, the instrument, and the observed explanatory variables are informative about the distribution of the unobservables.

In closing, we caution against the potential for misuse of the idea of using observables to draw inferences about selection bias, whether through an informal comparison of means or through the estimators we propose. The conditions required for Theorem 1 imply that it is dangerous to infer too much about selection on the unobservables from selection on the observables if the observables are small in number and explanatory power, or if they are unlikely to be representative of the full range of factors that determine an outcome.
References


