Finding Exogenous Variation in Data

Elliot Abrams  
Booth School of Business, University of Chicago  
eabrams@uchicago.edu

George Gui, Ali Hortacşu  
University of Chicago  
{georgeui, hortacsu}@uchicago.edu

Abstract

We reconsider the classic problem of recovering exogenous variation from an endogenous regressor. Two-stage least squares recovers the exogenous variation through presuming the existence of an instrumental variable. We instead rely on the assumption that the regressor is a mixture of exogenous and endogenous observations—say as the result of a temporary natural experiment. With this assumption, we propose an alternative two-stage method based on nonparametrically estimating a mixture model to recover a subset of the exogenous observations. We demonstrate that our method recovers exogenous observations in simulation and can be used to find pricing experiments hidden in grocery store scanner data.

1 Introduction

Consider the classic instrumental variables setup in which \( Y = \alpha + \beta X + \epsilon, \text{Cov}(X, \epsilon) \neq 0 \), and there exists a \( Z \) such that \( \text{Cov}(Z, \epsilon) = 0 \) and \( \text{Cov}(Z, X) \neq 0 \). The problem faced here is that \( X \) contains a mix of both exogenous variation that can be used to identify \( \beta \) and endogenous variation that complicates this identification. The instrument \( Z \) allows the econometrician to isolate the exogenous variation. In the two-stage least squares (2SLS) solution, the econometrician runs the first stage \( X = \pi_0 + \pi_1 Z + u \) to recover the exogenous variation \( \tilde{X} = \pi_0 + \pi_1 Z \). The econometrician then uses the recovered exogenous variation to find a consistent estimate for \( \beta \) in a second stage.

The main roadblock to applying the 2SLS solution generally is finding a suitable \( Z \). That said, the 2SLS solution conceptually applies quite broadly. As a prototypical example, consider a widget store choosing prices for its widgets. The econometrician may see prices and quantities sold when \( X \) is a mixture model with exogenous and endogenous components, \( \epsilon_k \), and \( \epsilon_{kt} \). Specifying the joint distribution of \( Z \), \( X \), and \( \epsilon \) would facilitate parametric identification in our setting following the authors’ approach.

Motivated by such settings, we revisit the problem of extracting exogenous variation with a new approach. We show that when \( X \) is a mixture model with exogenous and endogenous components, we can use nonparametric estimation to recover a subset of the exogenous observations. Just as with 2SLS, the recovered exogenous observations can then be used in a second stage to identify the parameters of interest.

Our work is an operationalization of the recent literature on nonparametric estimation of mixture models. (Hall and Zhou 2003) consider a mixture of two component distributions each with \( k \) independent coordinates and prove that nonparametric identification holds when \( k \geq 3 \). (Benaglia, Chauveau, and Hunter 2009) present an expectation maximization-like algorithm for nonparametrically estimating a finite mixture of \( m \) arbitrary component distributions with \( r \) independent coordinates when \( 2^r - 1 \geq mr + 1 \). (Bonhomme, Jochmans, and Robin 2016) derive asymptotic theory in the sub-case where the \( r \) coordinates are independent and identically distributed.

We also parallel research on the estimation of mixture linear regressions. (Bashir and Carter 2012) consider a model where there are \( k \) latent populations each satisfying a linear model \( Y_k = X_k' \beta_k + \epsilon_k \). Assuming that the errors \( \epsilon_k \sim N(0, \sigma_k) \) are independent across populations, the authors provide an expectation maximization algorithm to recover the class labels, \( \beta_k \), and \( \sigma_k \). Our setting differs from the above because we allow for a population \( k \) that has \( X_k \), endogenous with \( \epsilon_{kt} \), where \( \epsilon_{kt} \) is independent of \( X_k \) that is independent of \( X \).

Finally, our work extends a computer science literature on using machine learning to identify causal relations. (Jensen et al. 2008) propose a system for automatically identifying quasi-experimental designs from relational databases. (Grosse-Wentrup et al. 2016) develop an algorithm for inferring causal relations in brain imaging data. Most closely related, (Sharma, Hofman, and Watts 2016) consider settings where \( Y \) can be split into (1) a random variable \( Y_t \) that is causal by \( X \) and (2) a random variable \( Y_D \) that is independent of \( X \), and \( X \) is independent of \( Y_t \) if there are no confounding unobserved variables that cause both \( Y \) and \( X \). Here, the authors can estimate the causal impact of \( X \) on \( Y_t \) from subsets of the data where \( X \perp Y_D \). In contrast, our method applies for any \( Y \) and

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
instead proceeds by splitting \( X \) into exogenous and endogenous variation.

The rest of the paper is organized as follows. Section 2 presents our method. Section 3 demonstrates our method on simulated data. Section 4 takes our method to observed grocery store scanner data. Section 5 concludes.

## 2 Method

Consider a random vector \((Y, X, \epsilon)\) satisfying \( Y = \alpha + \beta X + \epsilon \) where \( \text{Cov}(X, \epsilon) \neq 0 \). 2SLS recovers exogenous variation from \( X \) through assuming the existence of an instrument \( Z \) satisfying \( \text{Cov}(Z, \epsilon) = 0 \) and \( \text{Cov}(Z, X) \neq 0 \). However, few applications feature an instrument that meets these conditions. As such, we propose a new method for recovering exogenous variation from \( X \) in certain settings that do not rely on having a valid instrument.

We are interested in settings typified by the widget store example above where \( X \) is a mix of observations from natural experiments and the “ordinary course of business.” Similar in spirit to instrumental variables, we rely on the existence of two additional variables, \( W_1 \) and \( W_2 \), that likewise differ between natural experiments and the “ordinary course of business.” Formally, we consider settings where Assumption 1 below holds.

### Assumption 1

\((Y, X, \epsilon, W_1, W_2)\) is a random vector satisfying \( Y = \alpha + \beta X + \epsilon \). The sub-vector \((X, W_1, W_2)\) is a two component mixture model with independent coordinates defined by the irreducible density function:

\[
  f_{X,W_1,W_2} = w f_{X_1} f_{W_1} + (1-w) f_{X_2} f_{W_2}
\]

Further for \( X_1 \sim F_{X_1} \) and \( X_2 \sim F_{X_2} \), \( \text{Cov}(X_1, \epsilon) = 0 \) and \( \text{Cov}(X_2, \epsilon) \neq 0 \).

This assumption places weaker conditions on \( W_1 \) and \( W_2 \) than 2SLS places on \( Z \). We do not require that \( W_1 \) and \( W_2 \) have non-zero covariance with \( X \) nor that they have zero covariance with \( \epsilon \).

Given this setup and \( T \) iid observations on \((Y, X, W_1, W_2)\), the econometrician can recover a positive measure of the observations of \( X \) that are drawn from \( X_1 \) if she additionally knows either (A) whether \( w \) is greater or less than 0.5\(^1\) or (B) how a moment of a coordinate of \((X, W_1, W_2)\) differs between the two mixture components, e.g. \( E(X_1) > E(X_2) \).

Our method is to:

1. Nonparametrically estimate the mixture model defining \((X, W_1, W_2)\). Identification of the density up to a permutation of the component labels follows from (Benaglia, Chauveau, and Hunter 2009)
2. Label the two components based on the assumed \( w \) or the known moment condition(s)
3. Label observations of \( X \) that are drawn from \( X_1 \) with probability greater than \( p \) as exogenous

---

\(^1\)That is, knows that the exogenous observations are more or less common

---

The econometrician can then use the subset of observations from step (3) to consistently estimate \( \beta \) via ordinary least squares in a second stage. Specifically, let \( \chi(p) \) be a subset of the set of observations of \((Y_i, X_i)\) where \( X_i \) is drawn from \( X_1 \) with probability greater than or equal to \( p \) under the true mixture model. Theorem 2.1 below proves that the ordinary least squares estimator for \( \beta \) using the observations in \( \chi(p) \) is consistent as \( T \to \infty \) and \( p \to 1 \) under minor additional assumptions (proof in Appendix A.1). We demonstrate with a simulation and application below that accurate labels are achievable in practice.

### Theorem 2.1

Let \( \chi(p) \) be a subset of the set of observations of \((Y_i, X_i)\) where \( X_i \) is drawn from \( X_1 \) with probability greater than or equal to \( p \) under the true mixture model. Assume (1) \( \chi(p) \) approaches infinity as \( T \) approaches infinity and (2) the \( \text{Cov}(\epsilon, X|X \in \chi(p)) \) and \( \text{Var}(X|X \in \chi(p)) \) are finite for all \( p \in [0, 1] \).

Then

\[
  \hat{\beta}_{\text{OLS}} \xrightarrow{p,T \to \infty} \frac{\text{Cov}(Y, X|X \in \chi(p))}{\text{Var}(X|X \in \chi(p))} \xrightarrow{p \to 1} \frac{\text{Cov}(Y_1, X_1|X_1 \in \chi(p))}{\text{Var}(X_1|X_1 \in \chi(p))} = \beta
\]

where \( Y_1 \) indicates that the data-generating process for \( Y \) is in terms of \( X_1 \) only.

Proof: In Appendix A.1

## 3 Simulation

A simulation provides an immediate test of our approach. Let \((Y, X, \epsilon, W_1, W_2)\) be a random vector satisfying Assumption 1. Namely, \((X, W_1, W_2)\) is drawn from a random vector \((X_1, W_{11}, W_{12})\) with probability \( w \) and from a random vector \((X_2, W_{21}, W_{22})\) else. Let each coordinate of \((X_1, W_{11}, W_{12})\) be independently distributed as \( U(0, 1) \) and each coordinate of \((X_2, W_{21}, W_{22})\) be independently distributed as \( U(0, 2) \). Finally, assume that \( \epsilon = X_2 + v \) with \( v \sim U(0, 1) \), \( w = 0.4 \), \( \alpha = 0 \), and \( \beta = 2 \).

Consider an econometrician tasked with using \( T = 2,000 \) iid observations on \((Y, X, W_1, W_2)\) to recover \( \beta \). She could try running ordinary least squares on the entire sample. However, for our realizations, the resulting point estimate is 2.9, which is significantly different from 2 at the 1% level. See Column 1 of Table 1. Further, neither \( W_1 \) nor \( W_2 \) are valid instruments for \( X \), so instrumental variables is of no help here.

Rather than tossing out the data, we recommend that the econometrician use our method to find a subset of the 2,000 observations of \( X \) that are drawn from \( X_1 \) and so exogenous. If the econometrician knows (or is willing to assume) that \( E(X_1) < E(X_2) \), she directly recovers densities for each coordinate of the two components. Labeling these components according to the moment condition, results in 902 observations on \( X \) that are 90% likely to have been drawn from \( X_1 \). The econometrician can then run ordinary least

---

\(^2\)A sufficient condition is that at least one of the coordinates has separate supports between the two components

---

54
stores assigned to “Hi-Lo” for a category raised prices, and stores assigned to “EDLP” for a category lowered prices. Given this background, we hope to recover all store-week-category treatment labels. We then use the store-weeks we label as “Hi-Lo” and “EDLP” to estimate consumer demand elasticities by category in a second stage.

For a given category, the Dominick’s setting maps into our framework above. Let \( X_{j,t} \) be the demeaned log price of product \( j \) from the category in store \( i \) during week \( t \). Then the \( P \) product price vector, \( (X_{j1,t}, \ldots, X_{P,t}) \), is a three component mixture model—the components being the respective price distributions under “Control,” “Hi-Lo,” and “EDLP” pricing. Our method immediately extends to such three component mixtures in that the nonparametric estimation from (Benaglia, Chauveau, and Hunter 2009) still applies. However, we note that practice is ahead of theory here as there are no established sufficient conditions for nonparametric identification of three component mixture models with independent but not identically distributed coordinates. See the discussion in Benaglia (2009). Despite the lack of formal identification, the results below suggest that we are nevertheless able to recover the three components. We label these components as “Control,” “Hi-Lo,” or “EDLP” according to the moment conditions \( E(\sum X_{j,t} | Hi-Lo) > E(\sum X_{j,t} | Control) > E(\sum X_{j,t} | EDLP) \).

Table 2 illustrates our success at recovering store-week-category treatment labels for January 1992 through December 1993. Complicating this assessment is that Hoch et al. conducted additional pricing experiments after the documented end of the “Study 1” we consider here. As such, a lower bound on our accuracy is the number of store-week-category observations we correctly label across the full two years assuming that store-week-category observations not documented in “Study 1” are “Control.” This metric is reported in Column 3. Our method is at least about 45% accurate for the majority of categories. A more precise assessment of our accuracy is the number of store-week-category observations we correctly label across the documented experiment timeframe. This metric is reported in Column 6. Our method is 80% accurate on average and is practically perfect for 5 of the 10 categories.

Graphically displaying the recovered exogenous variation provides additional insight into our method’s success at identifying store-week-category treatment labels. Figure 1 shows the average demeaned log price for Cereals-RTE over time in Dominick’s store 86 classified into “Control”, “Hi-Lo”, and “EDLP.” Panel 1 depicts the labels we predict for each week and Panel 2 depicts the documented labels for each week. Comparing the panels, it is immediate that the nonparametric estimation almost perfectly recovers both the weeks documented as “Control” and the weeks documented as “Hi-Lo.” Further, the nonparametric estimation provides additional insight into our method’s success at recovering exogenous economic data. To this end, we apply our method to recover pricing experiments from retail scanner data. Specifically, we consider the Dominick’s scanner dataset maintained by the Kilts Center for Marketing (Dominick’s is a now shuttered Chicago-area grocery store chain). Hoch, Dreze, and Purk (1994) conducted a number of pricing experiments with Dominick’s in the timeframe covered by this data. We show that our approach recovers Hoch et al.’s “Study 1” pricing experiments with high accuracy.

We allow ourselves few liberties in applying our method to the Dominick’s data, and so do not present a full description of Hoch et al. here. The limited background knowledge that we permit ourselves is that Hoch et al. experimented with how Dominick’s grocery stores priced products in selected categories for several weeks in 1992 and 1993. Each store-category pair was assigned one of three treatment levels for the duration of the experiment: “Control,” “Hi-Lo,” or “EDLP.” Stores assigned to “Control” for a category kept pricing products following the chain’s standard procedure, stores assigned to “Hi-Lo” for a category raised prices, and stores assigned to “EDLP” for a category lowered prices.

Table 1: Regression results from 2,000 observations of the random vector \((Y, X, \epsilon, W_1, W_2)\) generated as described in Section 3. Column (1) displays the coefficient estimated using OLS on the full sample. Column (2) displays the coefficient estimated from observations that are more than 90% likely to have been drawn from \( X_1 \).

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>( Y )</th>
<th>( Y_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>2.907***</td>
<td>(0.022)</td>
</tr>
<tr>
<td>( X_p )</td>
<td>1.932***</td>
<td>(0.101)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.514***</td>
<td>(0.018)</td>
</tr>
<tr>
<td></td>
<td>0.970***</td>
<td>(0.059)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,000</td>
<td>902</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.899</td>
<td>0.289</td>
</tr>
</tbody>
</table>

Note: \(^*p<0.1; \quad **p<0.05; \quad ***p<0.01\)

4Dominick’s split their stores into comparable groups based on socioeconomic zones. We only work with stores from the largest such zone. We also exclude stores that have no product purchases from the category in over 15% of the weeks observed. This subsetting leaves us with 27 stores.

3In the estimation, we hand select \( P = 20 \) products.
tion also appears to correctly predict that the store continued “Hi-Lo” pricing for Cereals-RTE for several weeks following the end of Hoch et al.’s “Stage 1” experiment. Similarly, Figure 2 shows the average demeaned log price for Analgesics over time in Dominick’s store 107 classified into “Control”, “Hi-Lo”, and “EDLP.” Here, the nonparametric estimation almost perfectly recovers both the weeks documented as “Control” along with the weeks documented as “EDLP.” The nonparametric estimation also appears to correctly predict that the store returned back to “Control” pricing following the end of Hoch et al.’s “Stage 1” experiment.

Figure 1: Average demeaned log price for Cereals-RTE over time in Dominick’s store 86 classified into “Control”, “Hi-Lo”, and “EDLP”

Figure 2: Average demeaned log price for Analgesics over time in Dominick’s store 107 classified into “Control”, “Hi-Lo”, and “EDLP”

As a final step, we can use the recovered exogenous variation in a second stage to estimate demand elasticities by category. We estimate the demand elasticity for a category by regressing log quantity on log price for all store-week-product observations from the category for store-weeks we predict as “Hi-Lo” and “EDLP.” As the supports for “Hi-Lo,” “EDLP,” and “Control” pricing are visually distinct, we have some confidence that the resulting coefficient estimates are consistent by Theorem 2.1. Table 3 reports Hoch et al.’s estimates in Column 3 and our estimates in Column 4. For almost every category, the two estimates are quite close. That said, we note that ordinary least squares fit over all of 1992 and 1993 (rather than the subset we identify) also provides similar estimates. See Column 5. Still, our method is consistently closer to Hoch et al.’s estimates than OLS with week fixed effects estimated on the entire sample–a method some econometricians might consider using in this setting. See Column 6.

5 Conclusion

Two-stage least squares recovers exogenous variation from an endogenous regressor using an instrumental variable. Unfortunately, econometricians rarely have access to a valid instrument. As such, we revisit this problem with a new approach. Our key insight is that if the regressor is a mixture of exogenous and endogenous variation, then nonparametrically estimating the underlying mixture model recovers a subset of the exogenous observations. These recovered observations can then be used in a second stage to identify the parameter of economic interest.

Our method applies to the prototypical example of a widget store choosing prices at which to sell its widgets each week. These prices are endogenous in the regression of log quantity on log price because they are set simultaneously with demand. Assuming that the widget store either purposely experiments with its pricing practices for some period of time or events arise that produce a natural experiment, then there is a set of weeks in which prices are exogenous. Our method enables the econometrician to recover a subset of these weeks if she knows either (A) whether experiment weeks are more or less common than non-experiment weeks or (B) how a moment of an observed variable differs between experiment and non-experiment weeks.

Our approach has promise. It recovers exogenous observations in simulation and can be used to find pricing experiments hidden in retail scanner data. That is, an econometrician given observations on \( Y, X, W_1, W_2 \) satisfying Assumption 1 can still recover \( \beta \) without either \( W_1 \) or \( W_2 \) being valid instruments for \( X \). In practice, an econometrician given scanner data from a grocery store chain covering weeks in which the chain ran a pricing experiment can recover the essential details of the experiment–how each store was assigned to price products in each category each week.

Additional promising applications for our method include any settings where an econometrician wishes to recover lost or natural experiments. Consider a web company that regularly A/B tests changes to its website. Our understanding is that specific A/B tests conducted by one project team are not always stored for perpetuity and widely accessible to other project teams within the company. In such settings, our method can be used by future project teams to recover the historical tests from data on customer usage of the website over time. Similarly, technical difficulties or other issues
are likely to produce additional random variation in website usage. Our method can potentially be used to discover these natural experiments too.

An immediate extension of our method exists in situations where the econometrician has partial information on experiments in the data. For example, the econometrician may know details of some experiments that occur within the data and wish to find additional similar experiments. In these cases, the known details can be used to semiparametrically estimate the underlying mixture model (in place of relying on nonparametric estimation).

There are many settings where econometricians currently struggle to recover exogenous variation. We hope that our approach will facilitate advances in at least some of these instances. We expect that lessons from applying nonparametric estimation techniques will be informative on additional approaches to identification in the future.

6 Acknowledgments

We thank Patrick Bajari, Stéphane Bonhomme, Jean-Pierre Dubé, and Jeremy Fox for invaluable comments. We also thank Alan Montgomery and Peter Rossi for providing background documentation on the experiments from Hoch, Dreze, and Park 1994 and the Kilts Center for Marketing for access to the Dominick’s scanner dataset.
Applying the law of total variance to the numerator gives

\[ Cov(Y, X|X \in \chi(p)) = \]
\[ E[Cov(Y, X|1_{exp}, X \in \chi(p))|X \in \chi(p)] + \]
\[ Cov(E[Y|1_{exp}, X \in \chi(p)], E[X|1_{exp}, X \in \chi(p)]|X \in \chi(p)) \]

Then

\[ E[Cov(Y, X|1_{exp}, X \in \chi(p))|X \in \chi(p)] = \]
\[ pCov(Y, X|1_{exp} = 1, X \in \chi(p)) + \]
\[ (1 - p)Cov(Y, X|1_{exp} = 0, X \in \chi(p)) = \]
\[ pCov(Y_1, X_1|X_1 \in \chi(p)) + (1 - p)Cov(Y_2, X_2|X_2 \in \chi(p)) \]

As \( p \to 1, \)
\[ pVar(X_1|X_1 \in \chi(p)) + (1 - p)Var(X_2|X_2 \in \chi(p)) \]
\[ \to Var(X_1|X_1 \in \chi(1)) \]
\[ Var(E[X|1_{exp}, X \in \chi(p)]|X \in \chi(p)) \]
\[ \to 0 \]
\[ pCov(Y_1, X_1|X_1 \in \chi(p)) + (1 - p)Cov(Y_2, X_2|X_2 \in \chi(p)) \]
\[ \to Cov(Y_1, X_1|X_1 \in \chi(1)) \]
\[ Cov(E[Y|1_{exp}, X \in \chi(p)], E[X|1_{exp}, X \in \chi(p)]|X \in \chi(p)) \]
\[ \to 0 \]

So

\[ \hat{\beta}_{OLS} \xrightarrow{p \to \infty} \frac{Cov(Y, X|X \in \chi(p))}{Var(X|X \in \chi(p))} \]
\[ \to_{p \to 1} \frac{Cov(Y_1, X_1|X_1 \in \chi(p))}{Var(X_1|X_1 \in \chi(1))} \]
\[ = \frac{Cov(\alpha + \beta X_1, X_1|X_1 \in \chi(1))}{Var(X_1|X_1 \in \chi(1))} \]
\[ = \beta \]

A.2 Graphical Representations

Figure 3: Instrumental variables assumes the existence of a random variable \( Z \) that covaries with \( X \) and is independent of unobserved confounders, \( U_Y \), that cause both \( X \) and \( Y \)

Figure 4: We assume the existence of random variables \( W_1 \) and \( W_2 \) such that \( (X, W_1, W_2) \) has a two component mixture distribution. In contrast to instrumental variables, \( W_1 \) and \( W_2 \) need not be independent of the unobserved confounders

References


