Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 1–23

High-Dimensional Methods and Inference on Structural and Treatment Effects[†]

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen

Data with a large number of variables relative to the sample size—"high-dimensional data"—are readily available and increasingly common in empirical economics. High-dimensional data arise through a combination of two phenomena.

First, the data may be inherently high dimensional in that many different characteristics per observation are available. For example, the US Census, the Current Population Survey, the Survey of Income and Program Participation, the National Longitudinal Survey of Youth, and the American Housing Survey collect information on hundreds of individual characteristics. Economists are also increasingly using scanner datasets that record transaction-level data for households across a wide range of products, or text data where counts of words in documents may be used as variables. In both of these latter examples, there may be thousands or tens of thousands of available variables per observation.

Second, even when the number of available variables is relatively small, researchers rarely know the exact functional form with which the small number of variables enters the model of interest. Researchers are thus faced with a large set of potential variables formed by different ways of interacting and transforming the underlying variables.

■ *Alexandre Belloni is Associate Professor of Decision Sciences, Fuqua School of Business, Duke University, Durham, North Carolina. Victor Chernozhukov is Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Christian Hansen is Wallace W. Booth Professor of Econometrics and Statistics, University of Chicago Booth School of Business, Chicago, Illinois. Their email addresses are abn5@duke.edu, vchern@mit.edu, and chansen1@chicagobooth.edu.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.28.2.1

doi=10.1257/jep.28.2.1

There are many statistical methods available for constructing prediction models in the presence of high-dimensional data; for example, see Hastie, Tibshirani, and Friedman (2009) for a review. These methods tend to do a good job at prediction, which is what they are designed for, but they can often lead to incorrect conclusions when inference about model parameters such as regression coefficients is the object of interest (Leeb and Pötscher 2008a, b).

The goal of this paper is to provide an overview of how innovations in "data mining" can be adapted and modified to provide high-quality inference about model parameters. These data mining methods are relevant for learning about economic parameters where they are motivated, for example, by a desire to control properly for confounding variables. Note that here we use the term "data mining" in a modern sense which denotes a principled search for "true" predictive power that guards against false discovery and overfitting, does not erroneously equate in-sample fit to out-of-sample predictive ability, and accurately accounts for using the same data to examine many different hypotheses or models.

The key concept underlying the analysis of high-dimensional data is that dimension reduction or "regularization" is necessary to draw meaningful conclusions. The need for regularization can easily be seen when one considers an example where AQ1 there are exactly as many variables (plus a constant), as there are observations. In this case, the ordinary least squares estimator will fit the data perfectly, returning an R^2 of one. However, using the estimated model is likely to result in very poor forecasting properties out-of-sample because the model estimated by least squares is overfit: the least-squares fit captures not only the signal about how predictor variables may be used to forecast the outcome, but also fits the noise that is present in the given sample, and is not useful for forming out-of-sample predictions. Producing a useful forecasting model in this simple case requires regularization; that is, the estimates must be constrained so that overfitting is avoided and useful out-of-sample forecasts can be obtained.

We begin with a discussion of "approximately sparse" regression models in high-dimensional data. These models are characterized by having many potential predictor/control variables of which only a few are important for predicting the outcome. The challenge in this case is to obtain good out-of-sample forecasts of outcome (and/or treatment) variables without assuming that the researcher knows which of the many available variables actually correspond to the important predictors. We then turn to the issue of model selection with high-dimensional data when the goal is learning about specific model parameters. We show how methods designed for forecasting in approximately sparse regression models can be used in this context. To illustrate these ideas, we apply them to examples from three papers in the empirical literature: estimating the effect of eminent domain on house prices, estimating the effect of abortion on crime, and estimating the effect of institutions on economic output. Our focus is not to rework these studies in any complete way but to show how one can bring high-dimensional analysis into this work and how this introduction can influence the findings of the analysis.

Approximately Sparse Regression Models

To fix ideas, suppose we are interested in forecasting outcome y_i with controls w_i according to the model

$$y_i = g(w_i) + \zeta_i,$$

where the expected value of the error terms ζ_i given w_i is equal to zero. Further, suppose we have a sample of $i = 1, \ldots, n$ independent observations. To avoid overfitting and produce useful out-of-sample forecasts, we will generally need to restrict or regularize the function $g(\cdot)$.

There are many regularization approaches that produce the needed dimension reduction. Perhaps the simplest and most widely applied approach is the researcher making an ad hoc decision. Typically, applied researchers assume that they need only a small number of controls, which are chosen based on economic intuition and a modeling framework. Moreover, the researcher often assumes the controls enter the model in a simple fashion, usually linearly, allowing for the usual set of simple transformations and forming a small number of interaction terms. This approach has intuitive appeal and at some level is unavoidable. A researcher will always have to start by imposing some dimension reduction. However, it does leave one wondering whether the correct variables and functional forms were chosen.

Nonparametric methods, such as traditional series/sieve expansions, are also available. In this framework, one assumes that the model depends only on a small number of variables in a smooth but potentially unknown way and then uses a series of transformations of these variables in estimation; for example, see Newey (1997) and Chen (2007). Practical implementation of a nonparametric estimator requires that the researcher has selected an initial set of variables and a pre-specified set of series terms containing transformations of these variables. While more flexible than parametrics, traditional nonparametrics has a number of important limitations. Most importantly, it is again assumed that the most important terms for predicting are contained within a pre-specified set of transformed variables determined by the researcher that is quite small relative to the number of observations.

In this paper, we focus on an approach to regularization that treats $g(w_i)$ as a high-dimensional, approximately linear model.[1] Specifically, we assume that

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{j,i} + r_{p,i}.$$

[1] There is also work on high-dimensional nonlinear models. For example, van de Geer (2008) and Belloni, Chernozhukov, and Wei (2013) consider high-dimensional generalized linear regression, and Belloni and Chernozhukov (2011), Belloni, Chernozhukov, and Kato (2013), and Kato (2011) consider quantile regression. We consider only linear models here for simplicity. The basic insights from the high-dimensional linear models extend to nonlinear settings though the theoretical analysis and practical computation is more complicated. High-dimensional linear models can also encompass many interesting settings and can accommodate flexible functional approximation just as nonparametric series estimators can.

The variables $x_i = (x_{1,i}, \ldots, x_{p,i})'$ may simply be the elementary regressors w_i or may be made of transformations of these elementary regressors as in series modeling. In contrast to series modeling, we allow the number of these variables p to be larger than the sample size n. The final term $r_{p,i}$ is an approximation error. As with series, it is assumed that $r_{p,i}$ is small enough relative to sampling error in a well-defined sense (Bickel, Ritov, and Tsybakov 2009; Belloni, Chen, Chernohukov, and Hansen 2012). Without further restrictions on the model, practical inference in this kind of high-dimensional linear model remains impossible since $p \geq n$ is allowed.

A structure that has played an important role in the literature is *approximate sparsity* of the high-dimensional linear model. Approximate sparsity imposes a restriction that only s variables among all of $x_{j,i}$, where s is much smaller than n, have associated coefficients β_j that are different from 0, while permitting a nonzero approximation error $r_{p,i}$. Thus, estimators for this kind of model attempt to learn the identities of the variables with large nonzero coefficients, while simultaneously estimating these coefficients.[2]

Note that the approximately sparse high-dimensional linear model structure includes as a special case both the traditional parametric and nonparametric model. The approximately sparse high-dimensional model generalizes these approaches by allowing the researcher to consider many explanatory variables and to use the data to learn which of the many variables are the most important. This setting thus encompasses many usual approaches to data analysis and accommodates the realistic scenario where a researcher does not know a priori exactly which variables should be included in a model.

An appealing method for estimating the parameters of sparse high-dimensional linear models is the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Frank and Friedman (1993) and Tibshirani (1996), where coefficients are chosen to minimize the sum of the squared residuals plus a penalty term that penalizes the size of the model through the sum of absolute values of the coefficients. In our discussion and empirical examples, we use a variant of the LASSO estimator that we proposed in Belloni, Chen, Chernozkukov, and Hansen (2012) defined as

$$\hat{\beta} = \arg\min_{b} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{i,j}\, b_{j})^{2} + \lambda \sum_{j=1}^{p} |b_j| \gamma_j,$$

where $\lambda > 0$ is the "penalty level" and γ_j are the "penalty loadings." The penalty loadings are chosen to insure basic equivariance of coefficient estimates to rescaling of $x_{i,j}$ and can also be chosen to address heteroskedasticity, clustering, and non-normality in model errors. For more discussion, see Belloni, Chen, Chernozkukov, and Hansen (2012), Belloni, Chernozhukov, Hansen, and Kozbur (2014), and the online appendix for this paper at http://e-jep.org.

[2] Much of the high-dimensional linear model literature assumes that the model is exactly sparse, so the approximation error is identically 0. The approximately sparse model is strictly more general in that it allows for a nonzero approximation error in the analysis.

The penalty level, λ, controls the degree of penalization. Practical choices for λ that provably guard against overfitting are provided in Belloni, Chen, Chernozkukov, and Hansen (2012). (See also Belloni, Chernozhukov, Fernandéz-Val, and Hansen 2013; Belloni, Chernozhukov, Hansen, and Kozbur 2014.) It is also common to choose λ by cross-validation in prediction contexts, though it is important to note that this choice may not immediately equate to good performance when prediction is not the end goal.

The penalty function in the LASSO is special in that it has a kink at 0, which results in a sparse estimator with many coefficients set exactly to zero. Thus, the LASSO estimator may be used for variable selection by simply selecting the variables with nonzero estimated coefficients. A large part of the appeal of the LASSO estimator relative to other selection methods is that the LASSO problem is a convex optimization problem and highly efficient computational algorithms exist for its solution. LASSO-type estimators have also been shown to have appealing properties under plausible assumptions that allow for approximation errors, heteroskedasticity, clustering and fixed effects, and non-normality (Bickel, Ritov, and Tsybakov 2009; Belloni, Chen, Chernozhukov, and Hansen 2012; Belloni, Chernozhukov, Hansen, and Kozbur 2014; Gautier and Tsybakov 2011).

Finally, it is important to note that the nonzero coefficients that are part of the solution to the LASSO problem tend to be substantially biased towards zero. An appealing method to alleviate this bias is to employ the Post-LASSO estimator as in Belloni and Chernozhukov (2013) and Belloni, Chen, Chernozhukov, and Hansen (2012). The Post-LASSO estimator works in two steps. First, LASSO is applied to determine which variables can be dropped from the standpoint of prediction. Then, coefficients on the remaining variables are estimated via ordinary least squares regression using only the variables with nonzero first-step estimated coefficients. The Post-LASSO estimator is convenient to implement and, as we show in Belloni and Chernozhukov (2013) and Belloni, Chen, Chernozhukov, and Hansen (2012), works as well as and often better than LASSO in terms of rates of convergence and bias.

Model Selection When the Goal is Causal Inference

Using LASSO as a method for penalized estimation of the coefficients of a sparse linear model is useful for obtaining forecasting rules and for estimating which variables have a strong association to an outcome in a sparse framework. However, naively using the results obtained from such a procedure to draw inferences about model parameters can be problematic.

Part of the difficulty in drawing inferences after regularization or model selection is that these procedures are designed for forecasting, not for inference about model parameters. This observation suggests that more desirable inference properties may be obtained if one focuses on model selection over the predictive parts of the economic problem—the reduced forms and first-stages—rather than using model selection in the structural model directly.

The more difficult problem with doing inference following model selection is that model selection mistakes may occur. If one could be sure that the variable selector would always choose exactly all of the variables with nonzero coefficients, one could simply use the data to select this set of variables and then use the selected set coupled with any conventional procedure to do estimation and inference about parameters of interest. The validity of this approach is delicate because it relies on perfect model selection. Once one allows for the realistic scenario where some variables may have small but nonzero effects, it is likely there will be model selection mistakes in which such variables are not selected. The omission of such variables then generally contaminates estimation and inference results based on the selected set of variables. This problem is not restricted to the high-dimensional setting but is present even in low-dimensional settings when model selection is considered. This intuition is formally developed in Leeb and Pötscher (2008a, b). Because model selection mistakes seem inevitable in realistic settings, it is important to develop inference procedures that are robust to such mistakes.

An element in the provision of this robustness that has been employed recently AQ2 is to focus on a small set of parameters of interest over which no model selection will be done, leaving model selection or regularization to be done only over "nuisance" parts of the problem. The estimation for the main parameters is then carried out using estimating equations that are *orthogonal* or *immune* to small perturbations in the nuisance parts. In Belloni, Chen, Cheernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2013), we provide an approach that does this in a canonical instrumental variable model; see also Ng and Bai (2009). In Belloni, Chernozhukov, and Hansen (2013) and Belloni, Chernozhukov, and Hansen (forthcoming), we provide an approach for inference about coefficients in a partially linear model, or about average treatment effects in a heterogeneous treatment effects model with binary treatment; see also Farrell (2013). In addition to showing how to obtain valid inference following model selection in canonical econometric models, these papers develop basic intuition helpful in understanding how inference following regularization may be performed outside of these models. Thus, we outline the approaches of these papers below.

Providing formal results for doing inference about parameters following model selection for other models relevant in applied economics is a topic of ongoing research. For example, in Belloni, Chernozhukov, Fernandéz-Val, and Hansen 2013, we consider the estimation of heterogeneous treatment effects with endogenous receipt of treatment and present a general framework for econometric models where the orthogonality condition is explained in detail.

Inference with Selection among Many Instruments

Consider the linear instrumental variables model with potentially many instruments

$$y_i = \alpha d_i + \varepsilon_i$$

$$d_i = z_i' \Pi + r_i + v_i,$$

where $E[\varepsilon_i \mid z_i] = E[v_i \mid z_i, r_i] = 0$ but $E[\varepsilon_i v_i] \neq 0$, leading to endogeneity. In this setting, d_i is a scalar endogenous variable of interest, z_i is a p-dimensional vector of instruments where the number of instruments p may be much larger than the number of observations,[3] and r_i is an approximation error. Allowing for a small number of included exogenous variables is straightforward by defining the variables in the model as residuals after partialing these variables out, and we suppress this case for simplicity. The results in Belloni, Chen, Chernozhukov, and Hansen (2012) also allow for a nonscalar but finite-dimensional treatment vector.

One approach to estimation and inference about α in this context is to select a small number of instruments from z_i to use in a conventional two-stage least squares estimation. In Belloni, Chen, Chernozhukov, and Hansen (2012), we provide a set of formal conditions under which conventional inference from the two-stage least squares estimator based on instruments selected by LASSO or another variable selection procedure is valid for learning about the parameter of interest, α. The key features that allow this can be illustrated by noting that this model cleanly fits into the heuristic outline for doing valid inference after using high-dimensional methods provided above. The parameter of interest, α, is finite-dimensional and there is no selection over whether d_i will be included in the model. The variable selection component of the problem is limited to the first-stage equation relating the endogenous variable to the instruments, which is a pure predictive relationship. Finally, the structure of the problem is such that model selection mistakes in which a valid instrument with a small but nonzero coefficient is left out of the first-stage will not substantially affect the second-stage estimator of α as long as other instruments with large coefficients are selected. In other words, the second-stage instrumental variable estimate is *orthogonal* or *immune* to variable selection errors where instruments with small, nonzero coefficients are mistakenly excluded from estimation.

Inference with Selection among Many Controls

As a more complex example, consider a linear model where a treatment variable, d_i, is taken as exogenous after conditioning on control variables:

$$y_i = \alpha d_i + x_i' \theta_y + r_{yi} + \zeta_i,$$

where $E[\zeta_i \mid d_i, x_i, r_{yi}] = 0$, x_i is a p-dimensional vector of controls where $p \gg n$ is allowed, r_{yi} is an approximation error, and the parameter of interest is α, the effect of the treatment on the outcome.

Before turning to a procedure that provides high-quality estimates and inferential statements about α, it is useful to discuss some intuitive benchmarks that do

[3] In the instrumental variables setting, there are many papers that examine the properties of various instrumental variables estimators under many-instrument asymptotics where the number of instruments p is allowed to increase with the sample size n in such a way that $p < n$ and $p/n \rightarrow \rho < 1$; see, e.g. Bekker (1994), Chao and Swanson (2005), Hansen, Hausman, and Newey (2008), and Hausman, Newey, Woutersen, Chao, and Swanson (2012). These approaches do not apply when $p \geq n$ and tend to perform poorly when $p/n \approx 1$.

not work. Considering such cases builds intuition concerning features that must be guarded against in applying high-dimensional methods in this and related contexts.

One naive approach would attempt to select control variables by applying LASSO to the equation above, forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. One could then try to estimate and do inference about α by applying ordinary least squares with y_i as the outcome, and d_i and any selected control variables as regressors. The problem with this approach can be seen by noting that LASSO and any other high-dimensional modeling device targets *prediction,* not learning about specific model parameters. From the standpoint of prediction, any variable that is highly correlated to the treatment variable will tend to be dropped since including such a variable will tend not to add much predictive power for the outcome given that the treatment is already in the model. Of course, the exclusion of a variable that is highly correlated to the treatment will lead to substantial *omitted-variables bias* if the coefficient in θ_y associated with the variable is nonzero. Such omissions will happen routinely in any procedure that looks just at the equation above.

There are two problems with the above naive approach. First, it ignores a key component to understanding omitted-variables bias, the relationship between the treatment variable and the controls. To aid in learning about this relationship, we introduce an additional "reduced form" relation between the treatment and controls:

$$d_i = x_i' \theta_d + r_{di} + v_i,$$

where $E[v_i | x_i, r_{di}] = 0$. The other problem is that the naive approach is based on a "structural" model where the target is to learn the treatment effect given controls, not an equation representing a forecasting rule for y_i given d_i and x_i. It is thus useful to transform the first equation of this section to a reduced form, *predictive* equation by substituting the equation introduced for d_i into the "structural" equation yielding the reduced form system:

$$y_i = x_i' (\alpha\theta_d + \theta_y) + (\alpha r_{di} + r_{yi}) + (\alpha v_i + \zeta_i) = x_i'\pi + r_{ci} + \varepsilon_i$$

$$d_i = x_i' \theta_d + r_{di} + v_i,$$

where $E[\varepsilon_i | x_i, r_{ci}] = 0$, r_{ci} is a composite approximation error, and the second equation is the same as above. Both of these equations represent predictive relationships, which may be estimated using high-dimensional methods.

Before turning to the recommended procedure, let us mention the second set of naive procedures that use only one of these two equations for selection. The problem with working with only one of the above equations is that single-equation approaches rely on there being no errors in variable selection. To see the problem with such an approach, note that applying a variable selection method to say the first equation for forecasting y_i with x_i will tend to select variables with large entries in coefficient vector π but will tend to miss variables with moderately-sized coefficients. However, missing variables that have strong predictive power for d_i, namely variables with large coefficients in θ_d, may lead to substantive *omitted-variables bias* in the estimator of α if the coefficients on these variables in θ_y are moderately-sized. Intuitively, any such variable has a moderate direct effect on the outcome that will be incorrectly misattributed to the effect of the treatment when this variable is strongly related to the treatment and the variable is not included in the regression. Similarly, if one applied a variable selection method to only the second equation for predicting d_i, one would potentially miss variables that have moderate-sized coefficients in predicting d_i but large direct effects on y_i. Such an omission may again lead to non-negligible omitted-variables bias.

To guard against such model selection mistakes, it is important to consider *both* equations for selection: we apply variable selection methods to *each* of the two reduced form equations and then use all of the selected controls in estimation of α. Thus, variable selection is used to select a set of variables that are useful for predicting y_i, say x_{yi}, and a set of variables that are useful for predicting d_i, say x_{di}. We then estimate α by ordinary least squares regression of y_i on d_i and the *union* of the variables selected for predicting y_i and d_i, contained in x_{yi} and x_{di}. We thus make sure we use variables that are important for either of the two predictive relationships to guard against the omitted-variables bias, discussed above, when estimating α.

Using both variable selection steps immunizes the resulting procedure against the types of model selection mistakes discussed above for single-equation procedures. Specifically, using the variables selected in both reduced form equations ensures that any variables that have large effects in either the "structural" equation for y_i or the reduced form equation for d_i are included in the model. Any excluded variables are therefore at most mildly associated to y_i and d_i, which greatly limits the scope for omitted-variables bias. It is also noteworthy that the "double selection" procedure implicitly estimates the residuals and ε_i and v_i regresses the estimates of **AQ3** ε_i on the estimates of v_i to construct an estimator of α, thereby providing a selection analog of Robinson's (1988) method for estimating the parameters of a partially linear model.

In Belloni, Chernozhukov, and Hansen (2013, forthcoming), we provide formal conditions under which this "double selection" procedure will lead to valid inference about α even when selection mistakes are allowed, and provide substantial simulation evidence that the procedure works across a wide variety of models. Using both selection steps also enhances efficiency by finding variables that are strongly predictive of the outcome and may remove residual variance.[4]

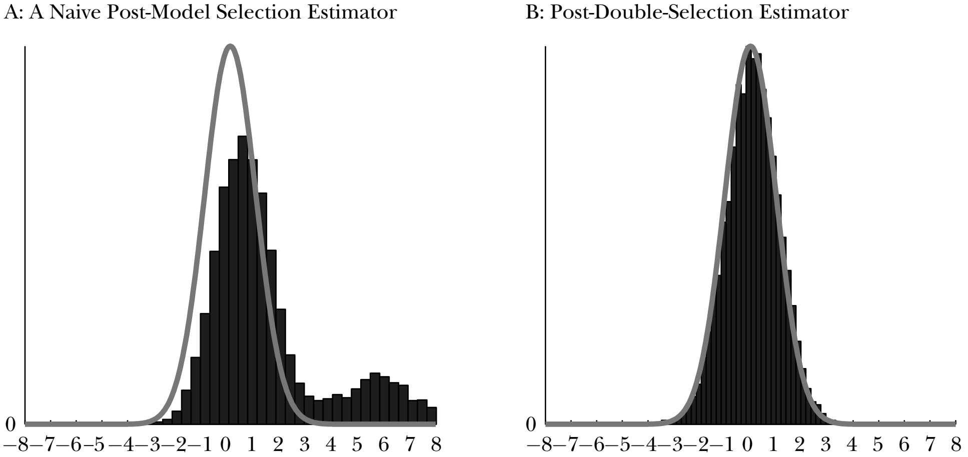
As a concrete illustration of the potential pitfalls of naive procedures and the robustness of the "double selection" approach, we present results from a simulation

[4] That is, standard errors may go down, at least theoretically, after performing the variable selection steps if the selected variables reduce residual variance sufficiently to offset the increased variability due to including more variables. In fact, under homoskedasticity, the estimator is semi-parametrically efficient, achieving the efficiency bound of Robinson (1988).

Figure 1

The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)

(distributions of estimators from each approach)


A: A Naive Post-Model Selection Estimator

B: Post-Double-Selection Estimator

Source: Belloni, Chernozhukov, and Hansen (forthcoming).

Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

F1 exercise in this linear modeling context in Figure 1. Details underlying the simulation are as in Belloni, Chernozhukov, and Hansen (2013). The left panel shows the sampling distribution of the estimator of α based on the first naive procedure discussed in this section, while the right panel shows the sampling distribution of the "double selection" estimator. The second mode in the left panel is due to model selection mistakes where important variables are missed leading to badly biased estimates of α. This strong omitted-variables bias is absent from the distribution of the "double selection" estimator, which was specifically designed to reduce the influence of such mistakes as discussed above.

Some Empirical Examples

In this section, we provide three concrete examples of the use of these methods. An online Appendix available with this paper at http://e-jep.org provides implementation details.

Estimating the Impact of Eminent Domain on House Prices

We consider instrumental variable estimation of the effects of federal appellate court decisions regarding eminent domain on housing prices. Recall that eminent domain refers to the government's taking of private property. Federal court rulings that a government seizure was unlawful (pro-plaintiff rulings) thus uphold individual property rights and make future exercise of eminent domain more difficult due to the structure of the US legal system. A more detailed discussion of the economics of takings law (or eminent domain) and other institutional and econometric considerations can be found in Belloni, Chen, Chernozhukov, and Hansen (2012) and Chen and Yeh (2012).

The analysis of the effects of takings law is complicated by the possible endogeneity between takings law decisions and economic variables: for example, a taking may be less likely if real estate prices are low and sellers are eager to unload property. To address the potential endogeneity of takings law, we employ an instrumental variables strategy based on the identification argument of Chen and Sethi (2010) and Chen and Yeh (2012) that relies on the random assignment of judges to federal appellate panels. Because judges are randomly assigned to three-judge panels to decide appellate cases, the exact identity of the judges and their demographics are randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year. Under this random assignment, the characteristics of judges serving on federal appellate panels can only be related to property prices through the judges' decisions; thus the judge's characteristics will plausibly satisfy the instrumental variable exclusion restriction.

Following this argument, we try to uncover the effect of takings law by estimating models of the form

$$\log(\mathit{Case\text{–}Shiller}_{ct}) = \alpha \cdot \mathit{TakingsLaw}_{ct} + \beta_c + \beta_t + \gamma_c t + W'_{ct}\delta + \varepsilon_{ct}$$

using the characteristics of judges actually assigned to cases as instruments for $\mathit{TakingsLaw}_{ct}$. In this equation, $\mathit{Case\text{–}Shiller}_{ct}$ is the average of the Case–Shiller home price index within circuit court c at time t; $\mathit{TakingsLaw}_{ct}$ represents the number of pro-plaintiff appellate takings decisions in federal circuit court c and year t; W_{ct} are included exogenous variables that include a dummy variable for whether there were relevant cases in that circuit-year, the number of takings appellate decisions, and controls for the distribution of characteristics of federal circuit court judges in a given circuit-year; and β_c, β_t, and $\gamma_c t$ are respectively circuit-specific effects, time-specific effects, and circuit-specific time trends. An appellate court decision is coded as pro-plaintiff if the court ruled that a taking was unlawful, thus overturning the government's seizure of the property in favor of the private owner. The parameter of interest, α, thus represents the effect of an additional decision upholding individual property rights on a measure of property prices. The sample size in this example is 183.

The argument given above suggests that judges' characteristics satisfy the instrumental variables exclusion restriction. Of course, to be valid instruments, the characteristics must also be useful for predicting judicial decisions. In the data, we observe a variety of demographic information about each judge, and the basic identification argument suggests that any set of characteristics of the three-judge panel will be unrelated to structural unobservables. Given the large number of instruments that could be constructed by considering all combinations of characteristics of three judge panels, it is also infeasible to just use all possible instruments.

Thus, a sensible way to proceed is to use variable selection methods to find a set of good instruments from a large set of intuitively chosen potential instruments. Under the exclusion restriction, the ideal set of instruments provides a high-quality prediction of the endogenous variable—judicial decisions in this example. Forming high-quality predictions, which is of course different from obtaining a good in-sample fit, is exactly what LASSO and other data mining procedures are designed to do. Note that using LASSO with proper penalty parameters theoretically guarantees that any instruments selected are not simply spuriously correlated to the endogenous variable but have true predictive power. This guarantee means that LASSO could select no instruments at all as there may be no set of variables with sufficient predictive power to achieve the required standard.

Intuitively, reliably distinguishing true predictive power from spurious association becomes more difficult as more variables are considered. This intuition can be seen in the theory of high-dimensional variable selection methods, and the methods work best in simulations when selection is done over a collection of variables that is not overly extensive. It is therefore important that some persuasive economic intuition exists to produce a carefully chosen, well-targeted set of variables to be selected over even when using automatic variable selection methods.

In this example, we first did dimension reduction by intuitively selecting characteristics thought to have strong signals about judge preferences over government versus individual property rights. We chose to consider only gender, race, religion (Jewish, Catholic, Protestant, evangelical, not-religious), party affiliation, source of academic degrees (bachelor's degree from an in-state university, bachelor's degree from a public university, JD from a public university, has an LLM or SJD), and whether the judge had been elevated from a district court. For each of these baseline variables, we then constructed three new variables, counting the number of panels with one member with each characteristic, two members with each characteristic, and three members with each characteristic. To allow for nonlinearities, we included first-order interactions between all of the previously mentioned variables, a cubic polynomial in the number of panels with at least one democrat, a cubic polynomial in the number of panels with at least one member with a JD from a public university, and a cubic polynomial in the number of panels with at least one member elevated from within the district. In addition to limiting the selection to be over this set of baseline variables, we did additional pre-processing to remove instruments that we thought likely to be irrelevant based on features of the instrument set alone. We removed any instrument where the standard deviation was extremely small and also removed one instrument from any pair of instruments that had a bivariate correlation exceeding .99 in absolute value. These instruments were removed as they were highly unlikely to have much power.[5] After these initial choices, we are left with a total of 147 instruments. The number of instruments plus the number of control variables is greater than the number of observations in this example, so conventional instrumental variables estimators using the full set of variables are not defined.

With this set of 147 instruments, we then estimate the first-stage relationship using LASSO. The estimated coefficients have just one nonzero element, the coefficient on the number of panels with one or more members with JD from a public university squared. Using this instrument gives a first-stage coefficient of 0.4495 with estimated standard error of 0.0511—that is, this variable appears to be a strong instrument. The second stage estimate using the LASSO-selected instrument is then 0.0648 with estimated standard error of 0.0240. This estimate is statistically significant at the usual levels, suggesting that a single additional judicial decision reinforcing individual property rights is associated with between 2 and 11 percent higher property prices with an average number of pro-plaintiff decisions per year of 0.19.

For comparison, we also experimented with an "intuitive" instrumental variable using the number of judicial panels with one or more Democrats. The political affiliation of judges is known to predict judicial decisions in several contexts, so one might hypothesize that this intuition carries over to judicial decisions regarding eminent domain. When we used the number of panels with one or more judges identified as Democrats as the single instrument, we found that one would not reject the hypothesis that this instrument is unrelated to the endogenous variable, the number of pro-plaintiff decisions, at any reasonable confidence level—which in turn suggests that this instrument is too weak to be useful.

We suspect that most analysts, relying on intuition about what might be a useful instrumental variable, would not have intuited that they should use the number of judicial panels with one or more members with JD from a public university squared. However, we find that one obtains a much stronger first-stage relationship in a two-stage least squares approach using this instrument selected by a formal variable selection method, relative to that obtained by an "intuitive" benchmark. This stronger first-stage in turn leads to a corresponding sensible and reasonably precise second-stage estimate. This substantive difference suggests that high-dimensional techniques may usefully complement researchers' intuition for choosing instruments and strengthen their ability to draw useful conclusions from the data.

Estimating the Effect of Legalized Abortion on Crime

Donohue and Levitt (2001) sought to estimate the effect of abortion on crime rates. Looking at state-level data, they find that higher rates of abortion in the years

[5] Note that selection based on characteristics of the instruments without reference to the endogenous variable or outcome cannot introduce bias as long as the instruments satisfy the instrumental variable exclusion restriction.

around 1970, as legal restrictions on abortion were eased in a number of states, are associated with lower rates of crime two decades later. However, there is a basic problem in estimating the causal impact of abortion on crime: state-level abortion rates during the earlier time period were not randomly assigned. It seems at least plausible that certain factors may be associated with both state-level abortion rates and state-level crime rates. Failing to control for these factors will then lead to omitted-variables bias in the estimated abortion effect.

To address these potential confounding factors, Donohue and Levitt (2001) estimate a differences-in-differences style model for state-level crime rates running from 1985 to 1997. Their basic specification is

$$y_{cit} = \alpha_c a_{cit} + w_{it}'\beta_c + \delta_{ci} + \gamma_{ct} + \varepsilon_{cit}.$$

The dependent variable, y_{cit}, indexes the crime-rate for crime type c (categorized between *violent, property,* and *murder*) in state i in year t. On the right-hand side, the independent variables are a_{cit}, a measure of the abortion rate relevant for type of crime c (as determined by the ages of criminals when they tend to commit crimes); w_{it}, a set of control variables to control for time-varying confounding state-level factors; δ_{ci}, state-specific effects that control for any time-invariant state-specific characteristics; and γ_{ct}, time-specific effects that control for national aggregate
AQ4 trends. For independent variables a_{cit}, Donohue and Levitt (2001) include the log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, the generosity of the Aid to Families with Dependent Children (AFDC) welfare program at time $t - 15$, a dummy for having a concealed weapons law, and beer consumption per capita as w_{it}. They present baseline estimation results based on this formulation as well as results from different models which vary the sample and set of controls in their tables IV and V. We refer the reader to the original paper for additional details, data definitions, and institutional background.

In this example, we take first-differences of the basic Donahue-Levitt formulation as our baseline. We use the same state-level data as Donohue and Levitt (2001) but delete Washington, DC,[6] which gives a sample with 50 cross-sectional observations and 12 time-series observations for a total of 600 observations. With these deletions, our baseline estimates using the same controls are quite similar to those reported in Donohue and Levitt (2001). Estimates of the effect of abortion
T1 on crime from this first-difference model are given in the first row of Table 1. These baseline results suggest that increases in abortion rates are strongly associated with decreases in crime rates; for example, an increase in the effective abortion rate of

[6] Removing Washington DC produces results similar to those in Donohue and Levitt (2001) without the need to introduce the weights used in Donohue and Levitt (2001) and is done for simplicity. This similarity between the weighted results and unweighted results excluding Washington DC is also discussed in Donohue and Levitt (2001).

Table 1
Effect of Abortion on Crime

	Type of crime					
	Violent		*Property*		*Murder*	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. error*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

Notes: This table reports results from estimating the effect of abortion on violent crime, property crime, and murder. The row labeled "First-difference" gives baseline first-difference estimates using the controls from Donohue and Levitt (2001). The row labeled "All controls" includes a broad set of controls meant to allow flexible trends that vary with state-level characteristics. The row labeled "Double selection" reports results based on the double selection method outlined in this paper and selecting among the variables used in the "All controls" results.

100 per 1,000 live births is associated with around a 15 percent reduction in violent crime. This association may be taken as causal under the assumption that all potential confounding factors not captured in w_{it} are either time-invariant or captured by a national trend.

Due to the inclusion of state and time effects, the baseline specification **AQ5** will control for any factors related to abortion and crime rates that are either time-invariant or vary only at the national level. While this formulation is fairly flexible, it produces valid estimates of the causal effect of abortion on crime rates *only if* time-varying state-specific factors that are correlated to both abortion and crime rates are captured by a small set of characteristics. An approach that is sometimes used to help alleviate such concerns is to include a set of state-specific linear time trends in the model to account for differences in state-specific trends that may be related to both the outcome and treatment variable of interest. However, this approach introduces many additional variables. Perhaps more importantly, the assumption of a linear state-specific trend is questionable in many circumstances as an approximation and certainly cannot capture the evolution of variables such as the crime rate or the abortion rate over any long time horizon.

Instead of using state-specific linear trends, we consider a generalization of the baseline model that allows for nonlinear trends interacted with observed state-specific characteristics and then use variable selection methods to find potentially important confounding variables. This approach allows us to consider quite flexible models without including so many additional variables that it becomes mechanically impossible to learn about the abortion effect. A key choice in using high-dimensional variable selection methods is the set of candidate variables to consider. For this example, our choice of these variables was motivated by our desire to accommodate a flexible trend that might offer a sensible model of the evolution of abortion or crime rates over a 12-year period. To accomplish this, we use the double-selection procedure outlined in the previous section with models of the form

$$\Delta y_{cit} = \alpha_c \Delta a_{cit} + z_{cit}'\beta_c + \tilde{\gamma}_{ct} + \Delta\varepsilon_{cit}$$

$$\Delta a_{cit} = z_{cit}'\Pi_c + \tilde{\kappa}_{ct} + \Delta v_{cit}.$$

In this formulation, $\Delta y_{cit} = y_{cit} - y_{cit-1}$ and Δa_{cit}, $\Delta\varepsilon_{cit}$, and Δv_{cit} are defined similarly; $\tilde{\gamma}_{ct}$ and $\tilde{\kappa}_{ct}$ are time effects; z_{cit} is a large set of controls; and we have introduced an equation for the abortion rate to make the relation to the earlier discussion clear. z_{itc} consists of 284 variables made up of the levels, differences, initial level, initial difference, and within-state average of the eight state-specific time-varying observables, the initial level and initial difference of the abortion rate relevant for crime type c, quadratics in each of the preceding variables, interactions of all the aforementioned variables with t and t^2, and the main effects t and t^2. This set of variables corresponds to a cubic trend for the level of the crime rate and abortion rate that is allowed to depend on observed state-level characteristics.

Because the set of variables we consider has fewer elements than there are observations, we can estimate the abortion effect after controlling for the full set of variables. Results from ordinary least squares regression of the differenced crime rate on the differenced abortion rate, a full set of time dummies, and the full set of variables in z_{itc} are given in the second row of Table 1. The estimated abortion effects are extremely imprecise with confidence intervals at the usual levels including implausibly large negative and implausibly large positive values for the abortion effect across all three outcomes. Of course, very few researchers would consider using 284 controls with only 600 observations because of exactly this issue.

The final row of Table 1 provides the estimated abortion effects based on the double-selection method of Belloni, Chernozhukov, and Hansen (forthcoming). At each stage of the process, we include the full set of time dummies without penalizing the parameters on these variables, which results in their selection in all cases, as we wish to allow for a flexible aggregate trend. In this example, we use LASSO to select variables from z_{cit} that are useful for predicting the change in crime rate Δy_{cit} and the change in the associated abortion rate. We then use the union of the set of selected variables, including time effects, as controls in a final ordinary least squares regression of Δy_{cit} on Δa_{cit}. In all equations, the selected variables suggest the presence of a nonlinear trend that depends on state-specific characteristics.[7] Looking at

[7] For violent crime, lagged prisoners, lagged police, lagged police × t, the initial income difference, the initial income difference × t, the initial beer consumption difference × t, average income, average income × t, and the initial abortion rate are selected in the abortion equation; and no variables are selected in the crime equation. For property crime, lagged prisoners, lagged police, lagged income, the initial income difference, the initial income difference × t, average income, and the initial abortion rate are selected in the abortion equation; and initial income squared × t^2 and average income squared × t^2 are selected in the crime equation. For the murder rate, lagged prisoners, lagged prisoners × t, lagged police × t, the initial income difference × t, average income × t, the initial abortion

the results, we see that estimated abortion effects are much more precise than the "kitchen sink" results that include all controls. However, the double-selection estimates for the effect of abortion on crime rates still produce 95 percent confidence intervals that encompass large positive and negative values.

It is interesting that one would draw qualitatively different conclusions from the estimates obtained using formal variable selection than from the estimates obtained using a small set of intuitively selected controls. Specifically, one would conclude that increases in abortion have a strong negative effect on crime rates using a small set of intuitively selected controls but would fail to reject the hypothesis that abortion is unrelated to crime rates at usual significance levels using estimates obtained using formal variable-selection. Of course, this comparison does not mean that the effect of the abortion rate provided in the first row of Table 1 is inaccurate for measuring the causal effect of abortion on crime. It does, however, imply that this conclusion is not robust to the presence of fairly parsimonious nonlinear trends. Foote and Goetz (2008) reach a similar conclusion based on an intuitive argument.[8]

Estimating the Effect of Institutions on Output

For our final example, we consider estimation of the effect of institutions on later levels of aggregate output following the work of Acemoglu, Johnson, and Robinson (2001). Estimating the effect of institutions on output is complicated by the clear potential for simultaneity between institutions and output: specifically, better institutions may lead to higher incomes, but higher incomes may also lead to the development of better institutions. To help overcome this simultaneity, Acemoglu, Johnson, and Robinson (2001) use mortality rates for early European settlers as an instrument for institution quality. The validity of this instrument hinges on the argument that settlers set up better institutions in places where they are more likely to establish long-term settlements; that where they are likely to settle for the long term is related to settler mortality at the time of initial colonization; and that institutions are highly persistent. The exclusion restriction for the instrumental variable is then motivated by the argument that GDP, while persistent, is unlikely to be strongly influenced by mortality in the previous century, or earlier, except through institutions.

In their paper, Acemoglu, Johnson, and Robinson (2001) note that their instrumental variable strategy will be invalidated if other factors are also highly persistent and related to the development of institutions within a country and to the country's GDP. A leading candidate for such a factor, as they discuss, is geography. Thus, they control for the distance from the equator in their baseline specifications and also

rate, and the initial abortion rate × t are selected in the abortion equation; and no variables are selected in the crime equation.

[8] See also Donohue and Levitt's (2008) response to Foote and Goetz (2008), which considers the same problem using a longer panel and finds similar results to Donohue and Levitt (2001) including state-specific linear time trends.

consider specifications with different sets of geographic controls such as dummy variables for continents; see their table 4.

As a complement to these results, we consider using high-dimensional methods to aid in estimating the model

$$\log(\mathit{GDPpercapita}_i) = \alpha \cdot \mathit{ProtectionfromExpropriation}_i + x_i'\beta + \varepsilon_i.$$

We use the same set of 64 country-level observations as Acemoglu, Johnson, and Robinson (2001). *ProtectionfromExpropriation* a measure of the strength of individual property rights that is used as a proxy for the strength of institutions, and x_i is a set of variables that are meant to control for geography. The underlying identifying assumption is the same as that employed in Acemoglu, Johnson, and Robinson (2001), which is that mortality risk is a valid instrument after controlling for geography. Acemoglu, Johnson, and Robinson (2001) address this by assuming that the confounding effect of geography is adequately captured by a linear term in distance from the equator or a set of dummy variables. The use of high-dimensional methods allow us to replace this assumption by the assumption that geography can be sufficiently controlled for by a small number of variables constructed from geographic information whose identities will be learned from the data.

To make use of high-dimensional methods, we note that the model in this example is equivalent to the three-equation system

$$\begin{aligned} \log(\mathit{GDPpercapita}_i) &= \alpha \cdot \mathit{ProtectionfromExpropriation}_i + x_i'\beta + \varepsilon_i \\ \mathit{ProtectionfromExpropriation}_i &= \pi_1 \cdot \mathit{SettlerMortality}_i + x_i'\Pi_2 + v_i \\ \mathit{SettlerMortality}_i &= x_i'\gamma + u_i, \end{aligned}$$

which yields three reduced form equations relating the structural variables to the controls:

$$\begin{aligned} \log(\mathit{GDPpercapita}_i) &= x_i'\tilde{\beta} + \tilde{\varepsilon}_i \\ \mathit{ProtectionfromExpropriation}_i &= x_i'\tilde{\Pi}_2 + \tilde{v}_i \\ \mathit{SettlerMortality}_i &= x_i'\gamma + u_i. \end{aligned}$$

We can thus select a set of control terms by carrying out variable selection for each of these three reduced form equations using the essential idea outlined in the discussion of selecting control variables. Valid estimation and inference for the parameter α can then proceed by conventional instrumental variable estimation using $\mathit{SettlerMortality}_i$ as an instrument for $\mathit{ProtectionfromExpropriation}_i$ with the union of variables selected from each reduced form as included control variables.

Table 2
Effect of Institutions on Output

	Latitude	*All controls*	*Double selection*
First stage	−0.5372 (0.1545)	−0.2182 (0.2011)	−0.5429 (0.1719)
Second stage	0.9692 (0.2128)	0.9891 (0.8005)	0.7710 (0.1971)

Notes: In an exercise that follows the work of Acemoglu, Johnson, and Robinson (2001), this table reports results from estimating the effect of institutions, using settler mortality as an instrument. The row "First Stage" gives the first-stage estimate of the coefficient on settler mortality obtained by regressing "$ProtectionfromExpropriation_i$" on "$SettlerMortality_i$" and the set of control variables indicated in the column heading. The row "Second stage" gives the estimate of the structural effect of institutions on log(GDP per capita) using "$SettlerMortality_i$" as the instrument and controlling for variables as indicated in the column heading (see text for details). Each column reports the results for different sets of control variables. The column "Latitude" controls linearly for distance from the equator. The column "All controls" includes 16 controls defined in the main text and in footnote 9, and the column "Double selection" uses the union of the set of controls selected by LASSO for predicting GDP per capita, for predicting institutions, and for predicting settler mortality. Standard errors are in parentheses.

It is important that a set of baseline variables be selected before variable selection methods are applied. Our target is to control for geography, so we consider a flexible but still parsimonious set of variables constructed from geography. Specifically, we set x_i equal to the dummy variables for Africa, Asia, North America, and South America plus a cubic-spline in latitude (altogether, twelve variables for latitude).[9]

T2 We report estimation results in Table 2. The first row of the table, labeled "First stage," gives the estimate of the coefficient on "*SettlerMortality*" from the first stage regression of "*ProtectionfromExpropriation*" on "*SettlerMortality*" and a set of control variables defined by the column headings, with the corresponding estimated standard error provided in parentheses below the coefficient estimate. The second row of the table, labeled "Second stage," gives the estimate of the structural effect of institutions on "log(*GDPpercapita*)" obtained by instrumental variables estimation of "log(*GDPpercapita*)" on "*ProtectionfromExpropriation*" using "*SettlerMortality*" as the instrument and controlling for variables as indicated in the column heading, with the estimated standard error again provided below the coefficient estimate in parentheses.

The first column of the table labeled "Latitude" gives baseline results that control linearly for latitude. These results correspond to the findings of Acemoglu, Johnson, and Robinson (2001), suggesting a strong positive effect of improved

[9] Specifically, we include latitude, latitude^2, latitude^3, $(\text{latitude}-.08)_+$, $(\text{latitude}-.16)_+$, $(\text{latitude}-.24)_+$, $((\text{latitude}-.08)_+)^2$, $((\text{latitude}-.16)_+)^2$, $((\text{latitude}-.24)_+)^2$, $((\text{latitude}-.08)_+)^3$, $((\text{latitude}-.16)_+)^3$, $((\text{latitude}-.24)_+)^3$ where latitude denotes the distance of a country from the equator normalized to be between 0 and 1, the breakpoints in the latitude function were chosen by taking round numbers near the quartiles of latitude, and $f(a) = (a)_+$ is shorthand notation for $f(a) = (a)1(a > 0)$ where $1(\cdot)$ is the indicator function that returns 1 when the expression inside the parentheses is true and 0 otherwise.

institutions on output with an underlying reasonably strong first-stage. This contrasts strongly with the second column of the table, which gives results controlling for all 16 of the variables defined above and in footnote 9. Controlling for the full set of terms results in a very imprecisely estimated first-stage. The estimate of the effect of institutions is then unreliable given the weak first-stage.

The variable selection methods discussed in this paper are constructed to produce a reasonable trade-off between this perhaps overly flexible second case and the parsimonious first case by allowing flexible functions to be considered but only using terms which are useful for understanding the underlying reduced form relationships. The final column of Table 2 labeled "Double selection" controls for the union of the set of variables selected by running LASSO on each of the three reduced form equations. The same single variable, the dummy variable for Africa, is selected in the reduced form equations for GDP and mortality, and no variables are selected in the reduced form for the expropriation variable. Thus, the final column is simply the conventional instrumental variable estimate with the Africa dummy included as the single control variable. The results are qualitatively similar to the baseline results, though the first-stage is somewhat weaker and the estimated effect of institutions is slightly attenuated though still very strong and positive. The slightly weaker first-stage suggests that the intuitive baseline obtained by controlling linearly for latitude may be inadequate, though the results are not substantively altered in this case. Again, we believe these results suggest that high-dimensional techniques may usefully complement the sets of sensitivity analyses that researchers are already doing (such as those underlying table 4 of Acemoglu, Johnson, and Robinson 2001). High-dimensional techniques can add rigor to these exercises and thus potentially strengthen the plausibility of conclusions drawn in applied economic papers.

Conclusion

The high-dimensional methods discussed in this paper provide a useful addition to the standard tools used in applied economic research. They allow researchers to perform inference about economically interesting model parameters in settings with rich confounding information. In these settings, dimension reduction is important if one hopes to learn from the data. We have emphasized variable selection methods for achieving this dimension reduction and outlined procedures that provide valid inference in simple, canonical models allowing for inevitable variable selection mistakes. In Belloni, Chernozhukov, Fernandéz-Val, and Hansen 2013, we provide an extended treatment showing that valid post-selection inference is generally available when estimation is based on orthogonal estimating equations.

We hope we have convinced you that data mining done correctly is the opposite of "bad practice": it an extremely useful tool that opens many doors in the analysis of interesting economic data. These tools allow researchers to add rigor and robustness to the "art" of variable or model selection in data analyses where the aim is to draw inferences about economically meaningful parameters. We have only skimmed the surface of the growing set of statistical methods appropriate to this setting and anticipate that there will be many developments that further extend these tools to cover an increasing set of economically relevant settings. Ultimately, the practice of considering high-dimensional data more openly coupled with appropriate methodology should strengthen the plausibility of inferences drawn from data and allow a deeper exploration of the structures underlying economic data.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91(5): 1369–1401.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 63(3): 657–81.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica* 80(6): 2369–2429.

Belloni, Alexandre, and Victor Chernozhukov. 2011. "ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models." *Annals of Statistics* 39(1): 82–130.

Belloni, Alexandre, and Victor Chernozhukov. 2013. "Least Squares after Model Selection in High-Dimensional Sparse Models." *Bernoulli* 19(2): 521–47.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernandéz-Val, and Christian Hansen. 2013. "Program Evaluation with High-dimensional Data." arXiv e-print, arXiv:1311.2645.

Belloni, Alexandre, Victor Chernozhukov, and Christian B. Hansen. 2013. "Inference for High-Dimensional Sparse Econometric Models." *Advances in Economics and Econometrics. 10th World Congress,* Vol. 3, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 245–95. Cambridge University Press.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. Forthcoming. "Inference on Treatment Effects after Selection amongst High-Dimensional Controls." *Review of Economic Studies.*

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2014. "Inference in High-Dimensional Panel Data." Unpublished paper; available from the authors by request.

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato. 2013. "Uniform Post Selection Inference for LAD Regression and Other Z-Estimation Problems." arXiv e-print, arXiv:1304.0282.

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei. 2013. "Honest Confidence Regions for Logistic Regression with a Large Number of Controls." arXiv e-print arXiv:1304.3969.

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics* 37(4): 1705–32.

Chao, John C., and Norman R. Swanson. 2005. "Consistent Estimation with a Large Number of Weak Instruments." *Econometrica* 73(5): 1673–92.

Chen, Daniel L., and Jasmin Sethi. 2010. "Insiders and Outsiders: Does Forbidding Sexual Harassment Exacerbate Gender Inequality?" Available at SSRN: http://ssrn.com/abstract=1119968.

Chen, Daniel L., and Susan Yeh. 2012. "Growth under the Shadow of Expropriation? The Economic Impacts of Eminent Domain." http://www.sole-jole.org/13463.pdf.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." Chap. 76 in *Handbook of Econometrics,* Vol. 6B, edited by James J. Heckman and Edward E. Leamer, 5559–5632. Elsevier.

Donohue, John J., III, and Steven D. Levitt. 2001. "The Impact of Legalized Abortion on Crime." *Quarterly Journal of Economics* 116(2): 379–420.

Donohue, John J., III, and Steven D. Levitt. 2008. "Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz." *Quarterly Journal of Economics* 123(1): 425–40.

Farrell, Max H. 2013. “Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations.” arXiv e-print, arXiv:1309.4686.

Foote, Christopher L., and Christopher F. Goetz. 2008. “The Impact of Legalized Abortion on Crime: Comment.” *Quarterly Journal of Economics* 123(1): 407–23.

Frank, Ildiko E., and Jerome H. Friedman. 1993. “A Statistical View of Some Chemometrics Regression Tools.” *Technometrics* 35(2): 109–135.

Gautier, Eric, and Alexander B. Tsybakov. 2011. “High-Dimensional Instrumental Variables Regression and Confidence Sets.” Available at arXiv.

Hansen, Christian, Jerry Hausman, and Whitney K. Newey. 2008. “Estimation with Many Instrumental Variables.” *Journal of Business and Economic Statistics* 26(4): 398–422.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer: New York, NY.

Hausman, Jerry, Whitney K. Newey, Tiemen Woutersen, John Chao, and Norman Swanson. 2012. “Instrumental Variable Estimation with Heteroskedasticity and Many Instruments.” *Quantitative Economics* 3(2): 211–55.

Kato, Kengo. 2011. “Group Lasso for High Dimensional Sparse Quantile Regression Models.” arXiv e-print, arXiv:1103.1458.

Leeb, Hannes, and Benedikt M. Pötscher. 2008a. “Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?” *Econometric Theory* 24(2): 338–76.

Leeb, Hannes and Benedikt M. Pötscher. 2008b. “Recent Developments in Model Selection and Related Areas.” *Econometric Theory* 24(2): 319–22.

Newey, Whitney K. 1997. “Convergence Rates and Asymptotic Normality for Series Estimators.” *Journal of Econometrics* 79(1): 147–68.

Robinson, Peter M. 1988. “Root-*N*-Consistent Semiparametric Regression.” *Econometrica* 56(4): 931–54.

Serena Ng, and Jushan Bai. 2009. “Selecting Instrumental Variables in a Data Rich Environment.” *Journal of Time Series Econometrics* 1(1).

Tibshirani, Rob. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society Series B* 58(1): 267–88.

van de Geer, Sara A. 2008. “High-Dimensional Generalized Linear Models and the LASSO.” *Annals of Statistics* 36(2): 614–45.

AUTHOR QUERIES

PLEASE ANSWER ALL AUTHOR QUERIES (numbered with “AQ” in the margin of the page). Please disregard all Editor Queries (numbered with “EQ” in the margins). They are reminders for the editorial staff.

AQ#	Question	Response
1.	I changed “, including a constant” to “(plus a constant)” Is that correct? A constant isn’t a variable, so I was confused.	
2.	You had “small dimensional set of parameters” I changed to “small set” (“low-dimensional” would be another possible solution, but not as good). OK?	
3.	You had “construct estimator of ” and I added the “an”. Correct? Just checking because it’s math so I’m not sure.	
4.	I added “For independent variables a^{cit}, “ before “Donohue and Levitt (2001).” Is that correct? If not where did they include these variables?	
5.	Please check this revision. The word “controls” sounded like a noun so I changed to “will control for”. Correct?	