cemmap
centre for microdata methods and practice

High-dimensional methods and inference on structural and treatment effects

Alexandre Belloni
Victor Chernozhukov
Chistian Hansen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP59/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

HIGH-DIMENSIONAL METHODS AND INFERENCE ON STRUCTURAL AND TREATMENT EFFECTS

A. BELLONI, V. CHERNOZHUKOV, AND C. HANSEN

The goal of many empirical papers in economics is to provide an estimate of the causal or structural effect of a change in a treatment or policy variable, such as a government intervention or a price, on another economically interesting variable, such as unemployment or amount of a product purchased. Applied economists attempting to estimate such structural effects face the problems that economically interesting quantities like government policies are rarely randomly assigned and that the available data are often high-dimensional. Failure to address either of these issues generally leads to incorrect inference about structural effects, so methodology that is appropriate for estimating and performing inference about these effects when treatment is not randomly assigned and there are many potential control variables provides a useful addition to the tools available to applied economists.

It is well-understood that naive application of forecasting methods does not yield valid inference about structural effects when treatment variables are not randomly assigned. The lack of random assignment of economic data has led to the adoption of estimation strategies among applied economists such as instrumental variables (IV) methods and conditional on observables estimators for treatment effects, the simplest of which is ordinary least squares (OLS) including control variables. However, these strategies are typically motivated and justified in a setting where the number of available variables is small relative to the sample size. Generally, these traditional estimators are ill-defined when the number of conditioning variables is greater than the sample size. Even in settings where the number of potential controls is smaller than the sample size, inference may be complicated due to standard approximations providing a poor guide to finite-sample behavior when the number of regressors is a non-vanishing fraction of the sample size, and many conventional estimators fail to even be consistent in this setting.[1] In very-high-dimensional settings where the number of controls may be larger than the sample size, estimation and informative inference using traditional methods is impossible; and some sort of dimension reduction is necessary if a researcher wants to learn from the data.

Date: First version: July 2, 2013. This version August 15, 2013.

[1]In the IV setting, there are many papers that examine the properties of various IV 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 $p/n \to \rho < 1$; see, e.g. Bekker (1994), Chao and Swanson (2005), Hansen, Hausman, and Newey (2008), and Hausman, Newey, Woutersen, Chao, and Swanson (2009). These approaches do not apply when $p \geq n$ and tend to perform poorly when $p/n \approx 1$.

High-dimensional data, data in which the number of variables is large relative to the sample size, is readily available and is becoming an increasingly common feature of data considered by applied researchers. High-dimensional data arise through a combination of two different phenomena. First, the data may be inherently high-dimensional in that data on many different characteristics of each observation are available. For example, many data sources that have been used for decades such as the SIPP, NLSY, CPS, U. S. Census, and American Housing Survey, to name a few, collect information on hundreds of individual characteristics. Economists are also increasingly using large databases such as scanner data-sets that record transaction level data for households across a wide-range of products or text data where counts of words or word combinations in documents may be used as variables. In both of these examples, there may be thousands or tens-of-thousands of available variables per observation. Second, even when the number of available variables is small, researchers who are willing to admit that they do not know the functional form with which the small number of variables enters the model of interest are faced with a large set of potential variables formed as interactions and transformations of the underlying variables. There are many statistical methods available for constructing forecasting models in the presence of high-dimensional data. However, it is well-known that these methods tend to do a good job at what they are designed for, forecasting, but often lead to incorrect conclusions when inference about model parameters such as regression coefficients is the object of interest; see Leeb and Pötscher (2008a; 2008b) and Pötscher (2009).

One way to address both lack of random assignment in many economically interesting problems and the presence of high-dimensional data is to note that traditional structural or treatment effects models may be recast as forecasting problems through the relevant first-stage or reduced form relationships. High-dimensional methods may then be used to construct good forecasting models for these reduced form quantities. Under some conditions, the structure gained from modelling these reduced form relationships may then be exploited to obtain estimates and valid inferential statements about structural effects of interest. The key condition which has been exploited thus far in the literature is that the reduced form relationships are approximately sparse; that is, good forecasts of the outcome and/or treatment variables may be obtained using a small number of the available controls whose identities are not known to the researcher but will be learned from the data. For example, see Bai and Ng (2009), Belloni, Chen, Chernozhukov, and Hansen (2012), and Gautier and Tsybakov (2011) for applications to IV model and Belloni, Chernozhukov, and Hansen (2011) for application to a linear model with many controls as well as models with heterogeneous treatment effects and a binary treatment variable.

1. Approximately Sparse Regression Models

A key concept underlying data-analysis with high-dimensional data is that dimension reduction (regularization) is necessary to draw meaningful conclusions. The need for regularization

can easily be seen when one considers an example where there exactly as many variables, including a constant, as there are observations. In this case, the OLS estimator perfectly fits the data, 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 over-fit: 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 but is not useful for forming out-of-sample predictions. Producing a useful forecasting model in this simple case requires regularization; i.e. estimates must be constrained so that overfitting is avoided and useful out-of-sample forecasts can be obtained.

1.1. **High-Dimensional-Linear-Models.** To fix ideas, suppose we are interested in forecasting outcome y_i with controls x_i according to the model

$$y_i = g(x_i) + \zeta_i \text{ where } \mathrm{E}[\zeta_i|x_i] = 0, \tag{1.1}$$

and that we have a sample of $i = 1, ..., n$ (independent) observations. Note that in writing down this model we have already imposed substantial regularization in that the function $g(\cdot)$ is not allowed to change arbitrarily across observations but is only allowed to differ for different values of x_i. It is also clear that further regularization is necessary when x_i may take on n values in the observed sample as any attempt to estimate the n values of $g(\cdot)$ at the observed values of x_i will perfectly fit the outcome without further restrictions.

Given its necessity, it is unsurprising that there are many available approaches to regularization. Perhaps the simplest and most widely applied approach in the context of estimation of treatment or structural effects is *ad hoc* dimension reduction by the researcher. Typically, an applied economist will assume that only a small number of controls is needed, the identities of which are chosen by the researcher using intuition, and that the controls enter the model in a simple fashion, usually linearly. I.e. the researcher assumes that $g(x_i) = z_i'\beta$ where z_i has $s \ll n$ elements and z_i may be made of the original x_i as well as transformations of this set of variables.[2] While this approach has intuitive appeal and at a deep level is unavoidable in that a researcher will always have to impose some *ex ante* dimension reduction driven by intuition, it does leave one wondering whether the correct variables and functional forms were chosen.

A related approach can be found in traditional nonparametrics using series or sieve expansions. In this approach, one assumes that a model depends only on a small number of variables in a smooth but potentially unknown way and then uses the first few terms in a basis expansion to approximate this relationship. Heuristically, one assumes that

$$g(x_i) = \sum_{j=1}^{s} \beta_j z_{j,i} + r_{s,i} \tag{1.2}$$

[2]It is common to relax the restriction that $\mathrm{E}[\zeta_i|x_i] = 0$ to $\mathrm{E}[z_i\zeta_i] = 0$ and estimate $y_i = z_i'\beta + \zeta_i$ where β is defined as the coefficients of the minimum mean-squared error linear approximation to $g(x_i)$.

where $\{z_{j,i} = p_j(x_i)\}_{j=1}^s$ are s approximating functions[3] formed from the x_i and $r_{s,i}$ is a remainder term. It is then assumed that the remainder term is uniformly small relative to sampling error so that its impact on the resulting estimator for $g(x_i)$ is negligible relative to sampling uncertainty and can thus be ignored. See, for example, Newey (1997), Chen (2007), or Chen and Pouzo (2009; 2012). As with the parametric model, practical implementation of a nonparametric estimator requires that the researcher has done some *ex ante* variable selection to come up with the initial set of variables x_i and the set of approximating functions. The number of basis terms to be used, s, will then typically be chosen based on cross-validation or an information criterion.[4] Relative to the fully parametric model, the nonparametric approach has the virtue of not completely specifying the form of $g(\cdot)$; and it is also closely related to the methods for high-dimensional-sparse-models that we recommend and outline below. Traditional nonparametrics are, however, still designed for an inherently low-dimensional setting where it is assumed that the most important terms for predicting y_i are contained within a pre-specified $s \ll n$ terms with the contribution of any remaining terms being negligible.[5]

A third approach to regularization is to model $g(x_i)$ as a high-dimensional-linear-model (HDLM).[6] In HDLMs, we assume that

$$g(x_i) = \sum_{j=1}^{p} \beta_j z_{j,i} + r_{p,i} \text{ where } p \gg n \tag{1.3}$$

is allowed and $r_{p,i}$ is a remainder term. The formulation is similar to that used for series-based nonparametric estimators except that we allow one to consider very many terms, $p \gg n$, in writing down the linear approximation. As with series, it is assumed that $r_{p,i}$ is uniformly small relative to sampling error.[7] Of course, without further restrictions on the model, practical inference in HDLMs is still impossible.

A structure which has gained a great deal of popularity is to assume that the HDLM is approximately sparse. Approximate sparsity imposes that only $s \ll n$ variables among the $p \gg n$ variables z_{ij} in (1.3) have associated coefficients β_j that are different from 0. Unlike

[3]Applied researchers often use simple power series where $p_j(x_i) = x_i^j$ in the scalar x_i case or dummy variable expansions where $p_j(x_i) = 1(x_i \in \mathcal{I}_j)$ and the support of x_i has been cut into s non-overlapping intervals $\mathcal{I}_1, ..., \mathcal{I}_s$.

[4]E.g. the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

[5]Theory typically assumes $s^2/n \to 0$ which approximates the case that the number of regressors considered may be moderate but should be a very small fraction of the available sample size.

[6]There is also work on high-dimensional-nonlinear-models. For example, van de Geer (2008), Bach (2010), and Belloni, Chernozhukov, and Wei (2013) consider high-dimensional logistic 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 HDLMs extend to nonlinear settings though the theoretical analysis and practical computation is more complicated. HDLMs also encompass many interesting settings and can accommodate flexible functional approximation just as nonparametric series estimators can.

[7]Note that since one could always choose the same approximating functions and number of terms as in the series case, the HDLM imposes a weaker set of conditions on the function $g(\cdot)$.

in the parametric or traditional nonparametric approach, it is not assumed that the identities of these s variables are known to the researcher *a priori*. The HDLM structure thus imposes that a good predictive model of the outcome may be formed by a model with many coefficients set exactly equal to zero. Thus, HDLM estimators that exploit the assumed structure involve a model selection component where one attempts to learn the identities of the variables with non-zero coefficients while simultaneously estimating these non-zero coefficients.

A simple and popular method for estimating the parameters of sparse HDLMs is bridge estimation where coefficients are chosen to minimize the sum of the squared residuals subject to a penalty. Formally, a conventional bridge estimator is defined as

$$\widehat{\beta}_r = \arg\min_b \sum_{i=1}^{n} (y_i - z_i'b)^2 + \lambda \sum_{j=1}^{p} |b_j|^r, \tag{1.4}$$

for $r > 0$.[8] The first term is the usual error sum of squares and would be minimized at the OLS estimator for β in the absence of the second term. The second term is a term that imposes a penalty for differences between the estimated coefficients and zero.[9] It is the presence of the second term that regularizes the problem and keeps the estimated model from perfectly fitting the sample in cases where $p \geq n$. The penalty level, λ, controls the degree of penalization and must be specified by the researcher much like a bandwidth in kernel estimation or a number of series terms in series estimation.

The parameter r controls the shape of the penalty term and is also specified by the researcher. Usual choices of r are $r = 2$ which corresponds to the well-known ridge regression and $r = 1$ which corresponds to the LASSO estimator introduced by Tibshirani (1996) and Frank and Friedman (1993). When $r > 1$, the solution obtained by solving (1.4) has coefficients shrunk towards zero but does not set any coefficients exactly equal to 0. For $0 < r \leq 1$, the penalty function is continuous but kinked (non-differentiable) at 0 which results in an estimator that will set some coefficients exactly to zero, thus providing variable selection, while simultaneously shrinking coefficients estimated to be non-zero towards 0. Another consideration when choosing r is that the penalty term is convex when $r \geq 1$, so $\widehat{\beta}_r$ is obtained by solving a convex optimization problem. For $0 < r < 1$, the optimization problem is no longer convex, and calculating $\widehat{\beta}_r$ may pose a computational hurdle. The LASSO estimator nicely straddles this boundary providing an estimator that does model selection while remaining an easily computable solution to a convex optimization problem.[10] LASSO-type estimators have

[8]To make estimates invariant to trivial rescaling of regressors, it is conventional to rescale each regressor to have sample second moment equal to one when using this formulation.

[9]Penalizing towards any fixed, known set of values is accomplished trivially by translating the coefficients.

[10]Many theoretical analyses of estimators for sparse HDLMs, including bridge estimators, focus on establishing so-called oracle results which show that HDLM estimators perform as well as the infeasible least squares estimator that regresses the outcome on just the s variables with non-zero coefficients; see Fan and Li (2001) or Zou (2006) for examples. Oracle properties can be established for bridge estimators with $r < 1$ as in Huang, Horowitz, and Ma (2008) as well as related estimators. However, establishing oracle properties relies on very stringent conditions including that the non-zero coefficients in the model are "far" from zero. In other

also been shown to have appealing properties under plausible assumptions that allow for approximation errors as in (1.3), heteroskedasticity, and non-normality; see Bickel, Ritov, and Tsybakov (2009) and Belloni, Chen, Chernozhukov, and Hansen (2012) among others. For these reasons, we focus on LASSO, (1.4) with $r = 1$, and variants throughout the remainder of this review.

Finally, it is important to note that the non-zero coefficients that are part of the solution to (1.4) tend to be substantially biased towards zero. For a large coefficient, this bias is more pronounced the larger the choice of r which partially motivates the use of penalized estimators with $r < 1$.[11] An intuitive alternative to using penalized estimators with $r < 1$ is to employ the Post-LASSO estimator as in Belloni and Chernozhukov (2013). The Post-LASSO estimator is simply conventional OLS regressing y on only the elements of z that are estimated to have non-zero coefficients in solving (1.4). With $r = 1$, the Post-LASSO estimator is extremely convenient to implement as it involves solving the LASSO problem for which fast algorithms exist in most statistical software packages and then running conventional OLS using a small number of variables. Belloni and Chernozhukov (2013) verify that this Post-LASSO estimator does no worse than the conventional LASSO and may substantially outperform it in many cases both theoretically and in simulations.

1.2. **Feasible LASSO Allowing Heteroskedasticity.** Heteroskedastic and non-Gaussian data is a common concern among applied economic researchers, and procedures that are robust to heteroskedasticity are routinely employed in empirical economics. Belloni, Chen, Chernozhukov, and Hansen (2012) provide a variant of the LASSO estimator and verify that it has good risk and model selection properties allowing for heteroskedastic and non-Gaussian data. They estimate the parameters of (1.3) by solving a weighted penalized optimization problem

$$\widehat{\beta}_L = \arg\min_b \sum_{i=1}^n (y_i - z_i'b)^2 + \lambda \sum_{j=1}^p |\widehat{\gamma}_j b_j| \tag{1.5}$$

where as before λ is the penalty level that controls the overall weight given to the penalty function and $\widehat{\gamma}_j$ are penalty loadings that help address heteroskedasticity and non-normality. As with the conventional LASSO, one can also obtain Post-LASSO estimates that reduce the shrinkage bias inherent in solving (1.5) by running conventional least squares using just the variables that were estimated to have non-zero coefficients:

$$\widetilde{\beta}_L = \arg\min_b \sum_{i=1}^n (y_i - z_i'b)^2 : \ b_j = 0, \ \text{ if } \ \widehat{\beta}_{Lj} = 0. \tag{1.6}$$

words, non-zero coefficients must be large enough to be distinguished from zero with very high probability in finite-samples which rules out variables with small but non-zero effects. This coefficient structure seems highly unrealistic in many economic applications, so we do not provide discussion of oracle results for HDLM estimators.

[11] It is less well-appreciated that there is also bias in estimates of small, but non-zero coefficients which is larger for $r < 1$. Generically, no choice of r can completely avoid these two sources of bias.

Belloni, Chen, Chernozhukov, and Hansen (2012) confirm that this Post-LASSO estimator continues to have good risk properties in heteroskedastic, non-Gaussian settings.

Implementation of (1.5) involves selection of penalty loadings $\widehat{\gamma}_j$ for $j = 1, ..., p$ and the penalty level λ. Rules-of-thumb for choosing λ are given in Belloni, Chen, Chernozhukov, and Hansen (2012). Simple choices are to set $\lambda = 2.2\sqrt{n}\Phi^{-1}(1 - \frac{q}{2p})$ for Φ^{-1} defined to be the quantile function of a standard normal random variable or $\lambda = 2.2\sqrt{2n\log(2p/q)}$ where $q \to 0$ is needed in either option in the theory.[12] In our examples, we set either $q = .05$ or $q = .1/\log(n)$.

Belloni, Chen, Chernozhukov, and Hansen (2012) show that ideally one would set the penalty weights equal to the infeasible values $\gamma_j = \sqrt{\frac{1}{n}\sum_{i=1}^n z_{ij}^2\zeta_i^2}$. These ideal weights are infeasible as they depend on the unknown ζ_i, but valid estimates can be obtained through a simple iterative scheme:

(1) For a fixed value of λ, choose an initial guess for b, say b_0, and calculate $\widehat{\zeta}_{i,0} = y_i - z_i'b_0$ for $i = 1, ..., n$. Use $\widehat{\zeta}_{i,0}$ to calculate $\widehat{\gamma}_{j,0} = \sqrt{\frac{1}{n}\sum_{i=1}^n z_{ij}^2\widehat{\zeta}_{i,0}^2}$ for $j = 1, ..., p$. Set $k = 1$.
(2) Solve (1.5) given λ and $\{\widehat{\gamma}_{j,k-1}\}_{j=1}^p$. Let b_k be the solution to (1.5) or the corresponding Post-LASSO estimates.
(3) Calculate $\widehat{\zeta}_{i,k} = y_i - z_i'b_k$ for $i = 1, ..., n$, and use $\widehat{\zeta}_{i,k}$ to calculate $\widehat{\gamma}_{j,k} = \sqrt{\frac{1}{n}\sum_{i=1}^n z_{ij}^2\widehat{\zeta}_{i,k}^2}$ for $j = 1, ..., p$.
(4) Stop if convergence is achieved or a maximum number of iterations is reached. Otherwise, set $k = k + 1$ and repeat steps (2)-(4).

Results in Belloni, Chen, Chernozhukov, and Hansen (2012) show that this procedure will be valid for estimating β. It is also worth noting that the theoretical results do not require that the procedure be iterated. However, simulations suggest that iteration improves the performance of the estimator and also that convergence is generally achieved after a relatively small number of iterations.

2. Model Selection Targeting Inference

In the preceding sections, we have outlined penalized estimation of the coefficients of a sparse linear model focusing on the use of LASSO. These methods are useful for obtaining forecasting rules for some outcome and for estimating which variables have a strong association to the outcome in the sparse framework. However, using the results obtained from such a procedure

[12] Another popular option is the use of cross-validation. Choosing λ by cross-validation is obviously computationally more demanding than using a simple plug-in rule-of-thumb. In addition, we found very similar results in our examples when taking λ as the largest value of λ which gives an average cross-validation error within one-standard deviation of the cross-validation minimizer as suggested in Friedman, Hastie, and Tibshirani (2010) using 10-fold cross-validation. Further exploration of the properties of cross-validation for selecting λ in this setting seems like an interesting avenue for additional research.

to perform inference about the values of the regression coefficients in (1.3) or to understand structural economic effects more generally is problematic.

The intuition for the difficulty in doing inference after regularization is clear in cases where one considers solutions to (1.4) with $r > 1$ and $p > n$. In this case, all variables enter the model with coefficients biased towards zero, and the bias of the individual coefficients is not estimable with $p > n$. Thus, valid inference about the population value of any individual coefficient is precluded in general. One might hope this problem is eliminated by imposing the sparse structure as only $s \ll n$ coefficients need to be learned. Indeed, within a sparse model, valid inference about non-zero coefficients following penalized estimation is possible if one believes that all non-zero coefficients are so large that they can essentially be perfectly distinguished from zero in finite-samples. Under this structure, one can perfectly learn the identities of the variables with non-zero coefficients and can thus use conventional methods to perform inference about the coefficients on these variables after learning their identities. However, the validity of this approach is very delicate as it relies on perfect model selection. Once one allows for variables that have moderately sized coefficients which cannot be perfectly differentiated from zero, there is a possibility of model selection mistakes in which such variables are not selected, and the omission of such variables then generally leads to a significant omitted variables bias.[13] This intuition is formally developed in Leeb and Pötscher (2008a; 2008b) and Pötscher (2009).

As a concrete illustration of this problem, we present results from a simple simulated example. Suppose that data are generated according to

$$y_i = d_i\alpha + x_i'\theta_g + \zeta_i, \quad \zeta_i \sim N(0,1) \tag{2.7}$$

$$d_i = x_i'\theta_m + v_i, \quad v_i \sim N(0,1) \tag{2.8}$$

where $\mathrm{E}[\zeta_i v_i] = 0$, $p = \dim(x_i) = 200$, the covariates $x_i \sim N(0,\Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$, $\alpha = .5$ is the parameter of interest, and the sample size n is set to 100. In this example, d_i represents a "treatment" variable whose effect conditional on the variables in x_i, α_0, we are interested in inferring. The coefficients on the control variables are set as $\theta_{g,j} = c_y\beta_j$ and $\theta_{m,j} = c_d\beta_j$ with $\beta_j = (1/j)^2$ for $j = 1, ..., 200$. The values of c_y and c_d are then chosen to generate population values for the R^2 of the (infeasible) regression of y onto x and the (infeasible) regression of d onto x of 0.5. This simulation model fits within the approximately sparse framework as the regression function in each equation is well-approximated using only the first few regressors and discarding the rest.

One approach to estimating α would be to estimate the parameters α and θ_g by applying LASSO to equation (2.7) without penalizing the coefficient on d to select a set of important controls from x, say x_c. One could then estimate α as the coefficient from the least squares regression of y onto d and the selected controls x_c, and inference about plausible values of α could proceed using the results from this regression. This approach would provide valid

[13] This problem is not restricted to the high-dimensional setting but is present even in low-dimensional settings when model selection is considered.

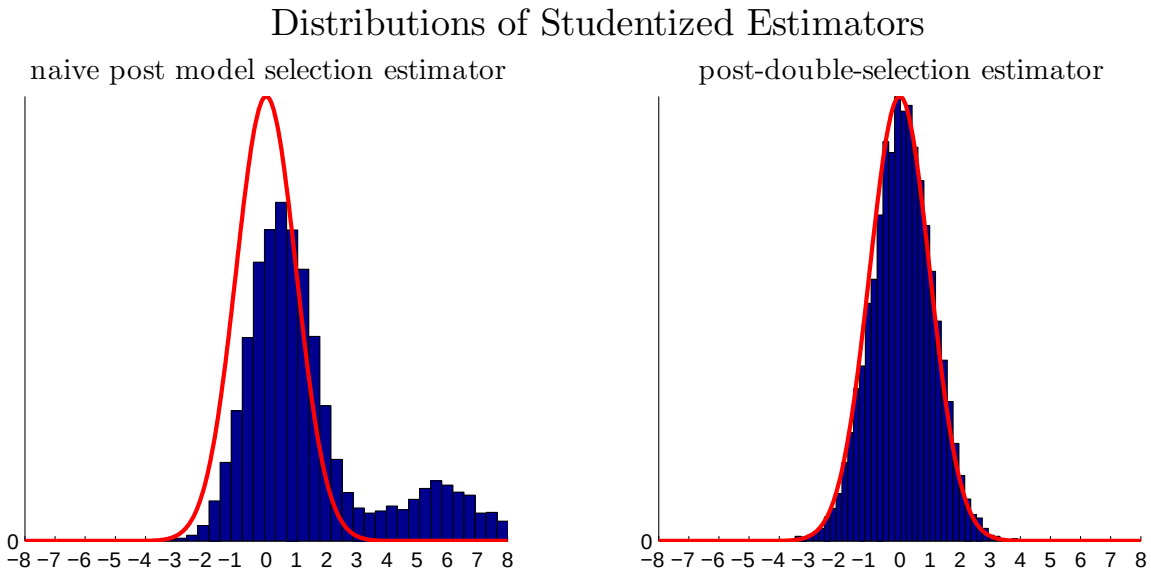

FIGURE 1. The finite-sample distributions (densities) of the naive post model selection estimator (left panel) and the post-model-selection estimator of Belloni, Chernozhukov, and Hansen (2012) designed targeting inference (right panel). The distributions are given for centered and studentized quantities.

inference if the model were truly sparse with a small number of elements in θ_g being large and the rest being identically equal to 0. The histogram in the left panel of Figure 1 displays the sampling distribution of the t-statistic based on this procedure and 10,000 simulation replications while the red curve gives the conventional asymptotic approximation. This figure illustrates the problem with naive post-model-selection inference, showing that the sampling distribution is bimodal and sharply deviates from the normal approximation. The second mode is driven by model selection mistakes in which variables with moderate coefficients are not included in the model.

Part of the difficulty in doing inference 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. Another key observation is that model selection mistakes are likely to occur in realistic settings where some variables may have small but non-zero partial effects; thus, it is important to develop inference procedures that are robust to such mistakes. An element to providing this robustness that has been employed recently is focusing on a small dimensional set of structural objects of interest over which no model selection will be done and leaving model selection or regularization only over "nuisance" parts of the problem. Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2013) provide an approach that does this in a canonical IV

model, and Belloni, Chernozhukov, and Hansen (2013; 2012) provide an approach for inference about structural coefficients in a partially linear model or about average treatment effects in a heterogeneous treatment effects model with binary treatment. In addition to being canonical econometric models, the basic intuition developed in these examples is helpful in understanding how regularization may be used in other environments, so we outline the approaches below. Developments in other contexts are a topic of ongoing research; e.g. Belloni, Chernozhukov, Fernandéz-Val, and Hansen (2013) consider estimating heterogeneous treatment effects in a program evaluation context.

2.1. **Inference with Selection among Many Instruments.** Consider the linear instrumental variables model with potentially many instruments

$$y_i = \alpha d_i + \varepsilon_i \tag{2.9}$$

$$d_i = z_i'\Pi + v_i \tag{2.10}$$

where $\mathrm{E}[\varepsilon_i|z_i] = \mathrm{E}[v_i|z_i] = 0$ but $\mathrm{E}[\varepsilon_i v_i] \neq 0$ leading to endogeneity, d_i is a scalar structural variable of interest, and z_i is a p dimensional vector of instruments where $p \gg n$ is allowed. Allowing for a small number of included exogenous variables is straightforward by defining the variables in (2.9) and (2.10) as residuals after partialing these variables out, and we ignore it for simplicity. The results in Belloni, Chen, Chernozhukov, and Hansen (2012) also allow for a non-scalar but finite dimensional treatment vector and for (2.10) to have the structure of (1.1).

One approach to estimation and inference about α in this context is to select a small number of instruments from z_i to then use in conventional 2SLS estimation. Belloni, Chen, Chernozhukov, and Hansen (2012) provide a set of formal conditions under which conventional inference from the 2SLS 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 also 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 (2.10) which is a pure predictive relationship. The structure of the problem is such that model selection mistakes in which a valid instrument with a small but non-zero coefficient is left out of the first-stage will not impact the consistency of the second-stage 2SLS estimator of α as long as there are other instruments with large coefficients that are selected. Such selection mistakes also do not have first-order impacts on the variance of the 2SLS estimator as they do not substantially diminish the performance of the first-stage *predictions* of d_i given z_i under the conditions that lead to model selection estimators leading to good predictive models. In other words, the second stage IV estimate is immunized against errors where variables with small coefficients are mistakenly excluded from estimation of the nuisance function $\mathrm{E}[d_i|z_i]$.

2.2. **Inference with Selection among Many Controls.** As a second example, consider a linear model where a treatment variable is taken as exogenous after conditioning on control variables:

$$y_i = \alpha d_i + x_i'\theta_g + \zeta_i \tag{2.11}$$

$$d_i = x_i'\theta_m + v_i \tag{2.12}$$

where $\mathrm{E}[\zeta_i|d_i, x_i] = \mathrm{E}[v_i|x_i] = 0$ and x_i is a p dimensional vector of controls where $p \gg n$ is allowed. Belloni, Chernozhukov, and Hansen (2012) consider this and more general models that allow for a non-scalar but finite dimensional treatment vector, for (2.11) and (2.12) to have the structure of (1.1), and for a binary treatment variable that is fully interacted with the controls allowing for general heterogenous effects and a binary treatment.

As was illustrated by the simulation example above, simply applying a variable selection procedure to (2.11) and then doing inference for α is complicated by the possibility of making selection errors. Rather than work directly with (2.11), it is more productive to consider the reduced forms associated with (2.11) and (2.12) formed by plugging the equation for the treatment (2.12) into the structural equation (2.11):

$$y_i = x_i'\pi + \varepsilon_i \tag{2.13}$$

$$d_i = x_i'\theta_m + v_i. \tag{2.14}$$

Both of these equations then represent simple predictive relationships which may be estimated using high-dimensional-methods. Interest focuses on the scalar parameter α, and there is no selection or shrinkage related to this parameter. Variable selection is limited to selecting a set of variables that are useful for predicting y, say x_y, and a set of variables that are useful for predicting d, say x_d. In order to estimate α, we then want to take the information obtained by building predictive models for y and d into account when estimating α from (2.11). A simple way to do this is to estimate equation (2.11) by OLS regression of y on d and the union of the variables selected for predicting y and d contained in x_y and x_d. This outlined procedure corresponds to the double-selection-method developed in Belloni, Chernozhukov, and Hansen (2012).

Belloni, Chernozhukov, and Hansen (2012) provide formal conditions under which this double-selection procedure will lead to valid inference about α even when selection mistakes are allowed in estimating both (2.13) and (2.14). The additional robustness relative to working with just the structural equation (2.11) comes from using the two selection steps and taking the union of the selected controls. If one took just the controls found to be highly predictive of d, i.e. used only x_d, one would potentially miss variables that have small but non-zero coefficients in (2.12) but large coefficients in (2.11). Ignoring these variables would then lead to non-negligible omitted variables bias. On the other hand, using only equation (2.11) or (2.13) would potentially miss variables that have small but non-zero coefficients in (2.11) but large coefficients in (2.12) which would again lead to non-negligible omitted variables bias. Using both variable selection steps immunizes the resulting procedure against both of these types of

model selection mistakes as the variables that are missed will have small coefficients in *both* (2.13) and (2.14) and will thus contribute little to the resulting omitted variables bias. Thus, the double-selection-procedure results in more robust inference than a procedure that relies on looking at only one equation or that does not take the union of the variables selected in both selection steps. Using both selection steps also enhances feasible efficiency by finding variables that are strongly predictive of the outcome and may remove residual variance.[14]

As a final illustration, we note that the right panel in Figure 1 results from applying this double-selection procedure in the simulation example presented at the beginning of this section. As can be seen, the sampling distribution of the t-statistic given by the histogram lines up quite nicely with the asymptotic distribution derived in Belloni, Chernozhukov, and Hansen (2012) which is represented by the red curve,[15] and inference based on the asymptotic approximation provides a good guide to the finite-sample performance of the double-selection estimator in this example.

3. Examples

In the preceding sections, we have briefly discussed regularization and variable selection through penalized estimation in linear models or series-based nonparametrics. We also provided two abstract examples illustrating how one may use sparse-high-dimensional methods coupled with econometric models to perform valid estimation and inference of structural parameters when there are many observed variables. In this section, we provide three concrete examples of the use of these methods as an aid to understanding economic phenomena. In all examples, variable selection is done using the feasible LASSO estimator outlined in Section 1.2 with penalty weights estimated via the iterative algorithm with a maximum of 100 iterations. All presented results also use a plug-in penalty level for λ.[16]

3.1. **Estimating the Impact of Eminent Domain on House Prices.**

In our first example, we consider IV estimation of the effects of federal appellate court decisions regarding eminent domain on the Case-Shiller Price Index. Our analysis is a simplified version of the analysis given in Belloni, Chen, Chernozhukov, and Hansen (2012) and Chen and Yeh (2010) which provide a more detailed discussion of the economics of takings law (or eminent domain), relevant institutional features of the legal system, and a careful discussion of endogeneity concerns and the instrumental variables strategy in this context. To try to uncover the relationship between

[14]Belloni, Chernozhukov, and Hansen (2012) show that this double-selection estimator achieves the semi-parametric efficiency bound when approximate sparsity as in (1.3) holds in both (2.13) and (2.14) and errors are homoskedastic.

[15]As the simulated object is the t-statistic using the standard error estimator from Belloni, Chernozhukov, and Hansen (2012), the red curve is a standard normal density.

[16]Using λ selected by 10-fold cross-validation yielded similar results in all cases.

takings law and housing prices, we estimate structural models of the form

$$\log(\text{Case-Shiller}_{ct}) = \alpha \text{Takings Law}_{ct} + \beta_c + \beta_t + \gamma_c t + W_{ct}'\delta + \epsilon_{ct}$$

where Case-Shiller_{ct} is the level of the Case-Shiller price index for circuit c at time t, Takings Law_{ct} represents the number of pro-plaintiff appellate takings decisions in circuit c and year t; W_{ct} are judicial pool characteristics,[17] a dummy for whether there were no cases in that circuit-year, and the number of takings appellate decisions; 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. We construe pro-plaintiff decisions to indicate a regime that is more protective of individual property rights. The parameter of interest, α, thus represents the effect of an additional decision upholding individual property rights on a measure of property prices. For simplicity and since all of the controls, instruments, and the endogenous variable vary only at the circuit-year level, we use the within-circuit-year average of the Case-Shiller index as the dependent variable. The total sample size in this example is 183.

The analysis of the effects of takings law is complicated by the possible endogeneity between takings law decisions and economic variables. 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 (2010) that relies on the random assignment of judges to federal appellate panels. Since 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.

There are many potential characteristics of three judge panels that may be used as instruments. While the basic identification argument suggests any set of characteristics of the three judge panel will be uncorrelated with the structural unobservable, there will clearly be some instruments which are more worthwhile than others in obtaining precise second-stage estimates. Given the large number of potential instruments that could be constructed by considering all combinations of characteristics of three judge panels, it is also clearly infeasible to use all possible instruments.

One approach to dealing with the large number of instruments is to choose a small number based solely on intuition. For example, judges' political affiliation is known to predict judicial decisions in several contexts. One might hypothesize that this intuition carries over to judicial decisions regarding eminent domain, perhaps with democratic leaning judges being more pro-government and thus tending to rule against the government's exercise of eminent domain less often. If one uses the number of panels with one or more democrats as the single instrument,

[17]The judicial pool characteristics are 32 variables for the probability of a panel being assigned with the characteristics used to construct the instruments.

the first-stage coefficient on the instrument is 0.0664 with an estimated standard error of 0.0713. Thus, 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 suggesting that this instrument does not satisfy the instrument relevance condition. Ignoring this, one can still obtain an IV estimate of the effect of an additional pro-plaintiff decision on property prices. Doing this gives a point estimate of -0.2583 with an estimated standard error of 0.5251, though one should not take these numbers too seriously given the obvious weak instrument.

An alternative would be to use variable selection methods to find a set of good instruments from a large set of intuitively chosen potential instruments. It is important to note that strong economic intuition is still needed even when using automatic variable selection methods as these methods will fail if the baseline set of variables being searched over is poor. It is also worth remembering that the theory of high-dimensional variable selection methods allows for selection among a very large set of variables but that high-dimensional variable selection methods work best in simulations when selection is done over a collection of variables that is not overly extensive. That is, it is important to have a carefully chosen, well-targeted set of variables to be selected over.

In this example, we first did *ex ante* dimension reduction by intuively selecting characteristics thought to have strong signal about judge preferences over government versus individual property rights. Specifically, we chose to consider only the individual characteristics gender, race, jewish, catholic, protestant, evangelical, not-religious, democrat, bachelor obtained in-state, bachelor from public university, JD from a public university, has an LLM or SJD, and whether 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 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 also did some 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 with mean $< .05$, any instrument with standard deviation after partialling out controls $< .000001$, and one instrument from any pair of instruments with bivariate correlation $> .99$ in absolute value.[18] After these initial choices, we are left with a total of 147 instruments.[19]

[18]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 IV exclusion restriction.

[19]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 instruments, we then estimate the first-stage relationship using LASSO, (1.5), with penalty level $\lambda = 2.2\sqrt{n}\Phi^{-1}(1-\gamma/(2p))$ where $\gamma = .1/\log(n)$. The estimated coefficients have just one non-zero 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. This strong first-stage is in sharp contrast to the first-stage obtained from the "intuitive" baseline using the number of panels with one or more democrats which had estimated coefficient (standard error) of 0.0664 (0.0713). The second stage estimate using the LASSO-selected instrument is then 0.0648 with estimated standard error of 0.0240. This estimate is small but statistically significant at the usual levels suggesting that judicial decisions reinforcing individual property rights are associated with higher property prices. That one obtains a much stronger first-stage using instruments selected by formal variable relative to that obtained by an "intuitive" benchmark with a corresponding sensible and reasonably precise second-stage estimate suggests that high-dimensional techniques may usefully complement researchers' intuition for choosing instruments in IV estimation settings.

3.2. **Estimating the Effect of Abortion on Crime.** As our second example, we consider estimating the impact of abortion on crime rates as in Donohue III and Levitt (2001). The basic problem in estimating the causal impact of abortion on crime is that state-level abortion rates are not randomly assigned, and it seems likely that there will be factors that are associated to 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 confounds, Donohue III and Levitt (2001) estimate a standard differences-in-differences style model for state-level crime rates running from 1985 to 1997. Their basic specification is

$$y_{cit} = \alpha_c a_{cit} + w_{it}'\beta_c + \delta_{ci} + \gamma_{ct} + \varepsilon_{cit} \tag{3.15}$$

where i indexes states, t indexes times, $c \in \{\text{violent, property, murder}\}$ indexes type of crime, δ_{ci} are state-specific effects that control for any time-invariant state-specific characteristics, γ_{ct} are time-specific effects that control for arbitrary national aggregate trends, w_{it} are a set of control variables to control for time-varying confounding state-level factors, a_{cit} is a measure of the abortion rate relevant for type of crime c,[20] and y_{cit} is the crime-rate for crime type c. Donohue III and Levitt (2001) use the log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, AFDC generosity at time $t-15$, a dummy for concealed weapons law, and beer consumption per capita as w_{it}. Tables IV and V in Donohue III and Levitt (2001) present baseline estimation results based on (3.15) as well as results from different models which vary the sample and set of controls to show

[20]This variable is constructed as weighted average of lagged abortion rates where weights are determined by the fraction of the type of crime committed by various age groups. For example, if all crime of type c were committed by 18 year olds, a_{cit} at time t would simply be the abortion rate at time $t-18$. See Donohue III and Levitt (2001) for further detail and exact construction methods.

TABLE 1. Effect of Abortion on Crime

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
First-Difference	-.152	.034	-.108	.022	-.204	.068
All Controls	.006	.755	-.154	.224	2.240	2.804
Double Selection	-.174	.120	-.052	.070	-.123	.148

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 III 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 Section 2.2 selecting among the variables used in the "All Controls" results.

that the baseline estimates are robust to small deviations from (3.15). We refer the reader to the original paper for additional details, data definitions, and institutional background.

In this example, we take first-differences of equation (3.15) as our baseline. We use the same state-level data as Donohue III and Levitt (2001) but delete Alaska, Hawaii, and Washington, D.C. which gives a sample with 48 cross-sectional observations and 12 time series observations for a total of 576 observations. With these deletions, our baseline estimates using the same controls as in (3.15) are quite similar to those reported in Donohue III and Levitt (2001). Estimates of the effect of abortion 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, and this association may be taken as causal under the assumption that all confounds are either time-invariant or captured by a national trend.

By construction, the baseline specification perfectly controls for any factors that are related to abortion and crime rates and either are time invariant or vary only at the national level due to the inclusion of the state and time effects. While this 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 the small set of characteristics given in x_{it}. An approach that is sometimes used to help alleviate this concern 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. This approach suffers from the drawback that it introduces many additional variables. Perhaps more importantly, the assumption of a linear 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 we mechanically cannot 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 Section 2.2 with models of the form

$$\begin{aligned} \Delta y_{cit} &= \alpha_c \Delta a_{cit} + z_{itc}'\beta_c + \widetilde{\gamma}_{ct} + \Delta\varepsilon_{cit} \\ \Delta a_{cit} &= z_{itc}'\Pi_c + \widetilde{\kappa}_{ct} + \Delta v_{cit} \end{aligned} \tag{3.16}$$

where $\Delta y_{cit} = y_{cit} - y_{cit-1}$ and Δa_{cit}, $\Delta\varepsilon_{cit}$, and Δv_{cit} are defined similarly; $\widetilde{\gamma}_{ct}$ and $\widetilde{\kappa}_{ct}$ are time effects; z_{itc} is a large set of controls; and we have introduced an equation for the abortion rate to make the relation to Section 2.2 clear. z_{itc} consists of 284 variables made up of the levels, differences, initial value, initial level, 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 flexible cubic trend for the level of the crime rate and abortion rate that is allowed to depend on observed state-level characteristics.

Since the set of variables we consider has fewere elements than there are observations, we can estimate the abortion effect after controlling for the full set of variables. Results from OLS 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. Unsurprisingly, 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 576 observations due to exactly this issue; and one may be willing to believe that many of the variables within this set of potential controls do not actually have any real association to the abortion rate or crime rate.

The final row of Table 1 provides the estimated abortion effects based on the double-selection method of Belloni, Chernozhukov, and Hansen (2012). At each stage of the process, we include the full set of time dummies without penalizing the parameters on these variables as we wish to allow for a flexible aggregate trend. In this example, we use LASSO with $\lambda = 2.2\sqrt{2n(\log(2p/.05))}$ to select variables from z_{cit} that are useful for predicting the change in crime rate c and the change in the associated abortion rate. We then use the union of the set of selected variables as controls in estimating (3.16). In all equations, the selected variables suggest the presence of a nonlinear trend in abortion rates that depends on state-specific

characteristics.[21] Looking at 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 are quite imprecise, producing 95% 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. Looking at the set of selected control variables, we see that the selected controls are strongly indicative of the presence of nonlinear trends that depend on state-specific characteristics. We also see that we cannot precisely determine the effect of the abortion rate on crime rates once one accounts for these trends. Of course, this 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. Interestingly, a similar conclusion is given in Foote and Goetz (2008) based on an intuitive argument.

3.3. **Estimating the Effect of Institutions on Output.** For our final example, we consider estimation of the effect of institutions on aggregate output following Acemoglu, Johnson, and Robinson (2001). Estimating the effect of institutions on output is complicated by the clear potential for simultaneity between institutions and output in that 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) make use of a clever instrumental variable strategy where they instrument for institution quality using early European settler mortality. The validity of this instrument hinges on the argument that settlers set up better institutions in places where they were more likely to establish long term settlements which is related to mortality at the time of initial colonization and that institutions are highly persistent which leads to a potential first-stage relationship. The exclusion restriction 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.

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

In their paper, Acemoglu, Johnson, and Robinson (2001) note that their IV strategy will be invalidated if there are other factors that are 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 that Acemoglu, Johnson, and Robinson (2001) discuss is geography. This possiblity leads to Acemoglu, Johnson, and Robinson (2001) controlling for the distance from the equator in their baseline specifications and also to their considering specifications with different sets of geographic controls such as continent dummies; see Acemoglu, Johnson, and Robinson (2001) Table 4.

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

$$\log(\text{GDP per capita}_i) = \alpha(\text{Protection from Expropriation}_i) + x_i'\beta + \varepsilon_i$$

using the same set of 64 country-level observations as Acemoglu, Johnson, and Robinson (2001) where "Protection from Expropriation" is 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) that mortality risk is a valid instrument after controlling for geography. Of course, as more flexible controls for geography are considered, power to identify the effect of institutions is lost. The specifications of Acemoglu, Johnson, and Robinson (2001) resolve this difficulty 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

$$\begin{aligned}
\log(\text{GDP per capita}_i) &= \alpha(\text{Protection from Expropriation}_i) + x_i'\beta + \varepsilon_i \\
\text{Protection from Expropriation}_i &= \pi_1(\text{Settler Mortality}_i) + x_i'\Pi_2 + v_i \\
\text{Settler Mortality}_i &= x_i'\gamma + u_i
\end{aligned}$$

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

$$\begin{aligned}
\log(\text{GDP per capita}_i) &= x_i'\widetilde{\beta} + \widetilde{\varepsilon}_i \\
\text{Protection from Expropriation}_i &= x_i'\widetilde{\Pi}_2 + \widetilde{v}_i \\
\text{Settler Mortality}_i &= x_i'\gamma + u_i.
\end{aligned}$$

After writing the model in terms of the reduced forms, the model is now seen to be structurally the same as the model from Section 2.2. We can thus select a set of control terms by performing variable selection on each of the three reduced form equations. Valid estimation and inference of the structural parameter, α, can then proceed by conventional IV estimation

TABLE 2. Effect of Institutions on Output

	Latitude	All Controls	Double Selection
First Stage	-0.5372	-0.2164	-0.5429
	(0.1545)	(0.2191)	(0.1719)
Second Stage	0.9692	0.9480	0.7710
	(0.2128)	(0.7384)	(0.1971)

This table reports results from estimating the effect of institutions using settler mortality as an instrument for different sets of control variables. The column labeled "Latitude" controls linearly for distance from the equator. The column labeled "All Controls" includes 16 controls defined in the text, and the column labeled "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. The row labeled "First Stage" gives the first stage estimate of the coefficient on settler mortality, and the row labeled "Second Stage" gives the estimate of the structural effect of institutions on log(GDP per capita). Standard errors are given in parentheses.

using Settler Mortality$_i$ as an instrument for Protection from Expropriation$_i$ with the union of variables selected from each reduced form as included control variables.

As in the previous examples, it is important that a set of baseline variables be selected *ex ante* before variable selection methods are applied. In this case, 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 the variables latitude, latitude^2, latitude^3, (latitude-.08)$_+$, (latitude-.16)$_+$, (latitude-.24)$_+$, ((latitude-.08)$_+$)^2, ((latitude-.16)$_+$)^2, ((latitude-.24)$_+$)^2, ((latitude-.08)$_+$)^3, ((latitude-.16)$_+$)^3, ((latitude-.24)$_+$)^3 where latitude denotes the distance of a country from the equator normalized to be between 0 and 1 and $(a)_+$ returns a when a is positive and 0 otherwise.

We report estimation results in Table 2. 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 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 in the previous paragraph. Controlling for the full set of terms results in a very imprecisely estimated first-stage. The effect of institutions is also very imprecisely estimated though the inference underlying this statement is unreliable given the weak first-stage. Of course, very few researchers would control for such a flexible function with only 64 available observations. The variable selection methods discussed in this paper are defined to produce a reasonable trade-off between this and the 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 (1.5) on each of the three reduced form equations with penalty level $\lambda = 2.2\sqrt{2n(\log(2p/\gamma))}$ with $\gamma = .1/\log(n)$. Interestingly, the same single variable, the dummy for Africa, is selected in all three of the reduced form equations. Thus, the final column is simply the IV estimate of the structural equation with the Africa dummy included as the single control variable. Interestingly, the results are qualitatively similar to the baseline results though the first-stage is somewhat weaker and the estimated structural effect 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) by adding rigor to these exercises and thus potentially strengthening the plausibility of conclusions drawn in applied economic papers.

4. Conclusion

In this paper, we consider estimation and inference in structural economic models allowing for very many conditioning variables or instruments. Researchers may face scenarios with many instruments or control variables in practice in settings where the data are inherently rich in that many characteristics about individual observations are available or in settings where researchers consider allowing variables to flexibly enter their models through the use of series or other expansions. Doing inference about structural effects is complicated in high-dimensional data as some sort of regularization or variable selection is necessary for informative inference to proceed. We briefly reviewed penalized methods for doing regularized estimation and variable selection in high-dimensional settings. While post-model-selection inference is generally complicated, we have illustrated through examples how valid inference may be performed after doing variable selection. In discussing the examples, we highlighted common features of the structural models that seem to be important in establishing that valid inference after model selection is possible. In three data examples, we have shown that high-dimensional methods may provide a useful addition to the tools used in applied economic research by allowing one to consider richer sorts of confounding information thus making arguments that rely on exogeneity of treatment or instrumental variables conditional on observables more plausible or by allowing one to search among a set of instruments to find those that lead to stronger identification. It is also worth noting that the methods discussed in this paper apply not only when there are very many regressors but may also be used when the number of regressors is smaller than sample size as in two of our three empirical examples.

References

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

Bach, F. (2010): "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, 4, 384–414.

Bai, J., and S. Ng (2009): "Selecting Instrumental Variables in a Data Rich Environment," *Journal of Time Series Econometrics*, 1(1).

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 63, 657–681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

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

——— (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19(2), 521–547.

Belloni, A., V. Chernozhukov, I. Fernandéz-Val, and C. Hansen (2013): "Program Evaluation with High-dimensional Data," *working paper*.

Belloni, A., V. Chernozhukov, and C. Hansen (2011): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *ArXiv*.

Belloni, A., V. Chernozhukov, and C. Hansen (2012): "Inference on treatment effects after selection amongst high-dimensional controls," .

Belloni, A., V. Chernozhukov, and C. Hansen (2013): "Inference for high-dimensional sparse econometric models," in *Advances in Economics and Econometrics. 10th World Congress of Econometric Society, August 2010.*, p. III:245295.

Belloni, A., V. Chernozhukov, and K. Kato (2013): "Uniform Post Selection Inference for LAD Regression Models," *arXiv preprint arXiv:1304.0282*.

Belloni, A., V. Chernozhukov, and Y. Wei (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," *arXiv preprint arXiv:1304.3969*.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, 37(4), 1705–1732.

Chao, J., and N. Swanson (2005): "Consistent Estimation With a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

Chen, D. L., and J. Sethi (2010): "Does Forbidding Sexual Harassment Exacerbate Gender Inequality," unpublished manuscript.

Chen, D. L., and S. Yeh (2010): "The Economic Impacts of Eminent Domain," unpublished manuscript.

Chen, X. (2007): "Large Sample Sieve Estimatin of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5559–5632.

Chen, X., and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152(1), 46–60.

——— (2012): "Estimation of nonparametric conditional moment models with possibly nonsmooth moments," *Econometrica*, 80(1), 277–322.

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

Fan, J., and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of American Statistical Association*, 96(456), 1348–1360.

Foote, C. L., and C. F. Goetz (2008): "The Impact of Legalized Abortion on Crime: Comment," *Quarterly Journal of Economics*, 123(1), 407–423.

Frank, I. E., and J. H. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35(2), 109–135.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33(1), 1–22.

Gautier, E., and A. B. Tsybakov (2011): "High-Dimensional Instrumental Variables Regression and Confidence Sets," *ArXiv working report*.

Hansen, C., J. Hausman, and W. K. Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26, 398–422.

Hausman, J., W. Newey, T. Woutersen, J. Chao, and N. Swanson (2009): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," mimeo.

Huang, J., J. L. Horowitz, and S. Ma (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *The Annals of Statistics*, 36(2), 587613.

Kato, K. (2011): "Group Lasso for high dimensional sparse quantile regression models," Preprint, ArXiv.

Leeb, H., and B. M. Pötscher (2008a): "Can one estimate the unconditional distribution of post-model-selection estimators?," *Econometric Theory*, 24(2), 338–376.

——— (2008b): "Recent developments in model selection and related areas," *Econometric Theory*, 24(2), 319–322.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Pötscher, B. (2009): "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhya*, 71-A, 1–18.

Tibshirani, R. (1996): "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

van de Geer, S. A. (2008): "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, 36(2), 614–645.

Zou, H. (2006): "The Adaptive Lasso And Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.