Journal of
Business &
Economic
Statistics

Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

Inference in High Dimensional Panel Models with an Application to Gun Control

Alexandre Belloni, Victor Chernozhukov, Christian Hansen & Damian Kozbur

INFERENCE IN HIGH DIMENSIONAL PANEL MODELS WITH AN APPLICATION TO GUN CONTROL

ALEXANDRE BELLONI, VICTOR CHERNOZHUKOV, CHRISTIAN HANSEN, AND DAMIAN KOZBUR

Abstract. We consider estimation and inference in panel data models with additive unobserved individual specific heterogeneity in a high dimensional setting. The setting allows the number of time varying regressors to be larger than the sample size. To make informative estimation and inference feasible, we require that the overall contribution of the time varying variables after eliminating the individual specific heterogeneity can be captured by a relatively small number of the available variables whose identities are unknown. This restriction allows the problem of estimation to proceed as a variable selection problem. Importantly, we treat the individual specific heterogeneity as fixed effects which allows this heterogeneity to be related to the observed time varying variables in an unspecified way and allows that this heterogeneity may be non-zero for all individuals. Within this framework, we provide procedures that give uniformly valid inference over a fixed subset of parameters in the canonical linear fixed effects model and over coefficients on a fixed vector of endogenous variables in panel data instrumental variables models with fixed effects and many instruments. We present simulation results in support of the theoretical developments and illustrate the use of the methods in an application aimed at estimating the effect of gun prevalence on crime rates.

Key Words: panel data, fixed effects, partially linear model, instrumental variables, high dimensional-sparse regression, inference under imperfect model selection, uniformly valid inference after model selection, clustered standard errors

1. Introduction

The use of panel data is extremely common in empirical economics. Panel data is appealing because it allows researchers to estimate the effects of variables of interest while accounting for time invariant individual specific heterogeneity in a flexible manner. For example, the widely used linear fixed effects model treats individual specific heterogeneity as a set of additive fixed effects to be estimated jointly with other model parameters and thus allows estimation of

Date: First version: April 2013. This version is of September 11, 2015.

We thank Stéphane Bonhomme, Elena Manresa, and seminar participants at UIUC, ETH Zurich, Toulouse School of Economics, Penn. State, Northwestern, UC Santa Barbara, and cemmap conference "Economic and Econometric Applications of Big Data" for helpful comments. We gratefully acknowledge financial support from the ETH Postdoctoral Fellowship and the NSF.

the common slope parameters of the model without imposing any structure over the additive individual specific heterogeneity.

Many panel data sets also have a large number of time varying variables available for each observation; i.e. they are "high dimensional" data. The large number of available variables may arise because the number of measured characteristics is large. For example, many panel data analyses in economics make use of county, state, or country level panels where there is a large set of measured characteristics and aggregates such as output, employment, demographic characteristics, etc. available for each observation. A large number of time varying variables may also be present due to a researcher wishing to allow for flexible dependence of an outcome variable on a small set of observed time varying covariates and thus considering a variety of transformations and interactions of the underlying set of variables. Identification of effects of interest in panel data contexts is also often achieved through a strategy where identification becomes more plausible as one allows for flexible trends that may differ across treatment states. Allowing for flexible trends that may differ based on observable characteristics may then be desirable but potentially introduces a large number of control variables.

A difficulty in high dimensional settings is that useful predictive models and informative inference about model parameters is complicated by the presence of the large number of explanatory variables. For example, the ordinary least squares estimator will fit the data perfectly if one uses a linear regression model in which there are exactly as many linearly independent explanatory variables as there are observations. However, the estimated model is likely to provide very poor out-of-sample predictions because the model estimated by unrestricted least squares is overfit. The least squares fit captures not just the signal about how the predictor variables may be used to forecast the outcome but also perfectly captures the noise in the given sample which is not useful for generating out-of-sample predictions. Constraining the estimated model to avoid perfectly fitting the sample data, or "regularization," is necessary for building a useful predictive model. Similarly, informative inference about parameters in a linear regression model is clearly impossible if the number of explanatory variables is larger than the sample size if one is unwilling to impose additional model structure.

A useful structure which has been employed in the recent econometrics literature focusing on inference in high dimensional settings is approximate sparsity; see, for example, Belloni, Chernozhukov, and Hansen (2010), Belloni, Chen, Chernozhukov, and Hansen (2012), and Belloni, Chernozhukov, and Hansen (2014). A leading example is the approximately sparse linear regression model which is characterized by having many covariates of which only a small number are important for predicting the outcome.Approximately sparse models nest conventional parametric regression models as well as standard sieve and series based nonparametric regression models. In addition to nesting standard econometric models, the framework is appealing as it reduces the problem of finding a good predictive model to a variable selection problem. Estimation methods appropriate for this framework also yield models with a relatively small set of variables which aids interpretability of the results and corresponds to the usual approach taken in empirical economics where models are typically estimated using a small number of control variables.

There are a variety of sensible variable selection estimators that are appropriate for estimating approximately sparse models. For example, ℓ_1-penalized methods such as the Lasso estimator of Frank and Friedman (1993) and Tibshirani (1996) have been proposed for model selection problems in high dimensional least squares problems in part because they are computationally efficient. Many ℓ_1-penalized methods and related methods have been shown to have good estimation properties with i.i.d. data even when perfect variable selection is not feasible; see, e.g., Candès and Tao (2007), Meinshausen and Yu (2009), Bickel, Ritov, and Tsybakov (2009), Huang, Horowitz, and Wei (2010), Belloni and Chernozhukov (2013) and the references therein. Such methods have also been shown to extend to nonparametric and non-Gaussian cases as in Bickel, Ritov, and Tsybakov (2009) and Belloni, Chen, Chernozhukov, and Hansen (2012), the latter of which also allows for conditional heteroscedasticity.

While the models and methods mentioned above are useful in a variety of contexts, they do not immediately apply to standard panel data models. There are two key points of departure between conventional approximately sparse high dimensional models and conventional panel data models used in empirical economics. The first is that the approximately sparse framework seems highly inappropriate for usual beliefs about individual specific heterogeneity in fixed effects models.[1] Specifically, the approximately sparse structure would imply that individual specific heterogeneity differs from some constant level for only a small number of individuals and may be completely ignored for the vast majority of individuals.

The second key difference is that the assumption of independent observations is inappropriate for many panel data sets used in economics. Many economic panels appear to exhibit substantial correlation between observations within the same cross-sectional unit of observation. It is well-known that failing to account for this correlation when doing inference about model parameters in panel data with a small number of covariates may lead to tests with substantial size distortions. This concern has led to the routine use of "clustered standard errors" which are robust to within-individual correlation and heterogeneity across individuals in empirical research; see Arellano (1987), Bertrand, Duflo, and Mullainathan (2004), and Hansen (2007) among others. In the context of variable selection in high dimensional models,

[1] There are other natural alternatives to dimension reduction over individual specific heterogeneity. See, e.g. Altonji and Matzkin (2005), Bester and Hansen (2009), Bester and Hansen (2014), and Bonhomme and Manresa (2013) for some recent examples.

failing to account for this correlation may result in substantial understatement of sampling variability. This understatement of sampling variability may then lead to a variable selection device selecting too many variables, many of which have no true association to the outcome of interest. The presence of these spuriously selected variables may have a substantial negative impact on the resulting estimator.

A key contribution of this paper is offering a variant of the Lasso estimator that accommodates a clustered covariance structure (Cluster-Lasso). We provide formal conditions under which the estimator performs well in the sense of returning a sparse estimate and having good forecasting and rate of convergence properties. By providing results allowing for a clustered error structure, we are also able to allow for the presence of unrestricted additive individual specific heterogeneity which are treated as fixed effects that are partialed out of the model before variable selection occurs. Accommodating this structure requires partialing out a number of covariates that is proportional to the sample size under some asymptotic sequences we consider. In general, partialing out a number of variables proportional to the sample size will induce a non-standard, potentially highly dependent covariance structure in the partialed-out data. The structure of the fixed effects model is such that partialing out the fixed effects cannot induce correlation across individuals, though it may induce strong correlation within the observations for each individual. Because this structure is already allowed for in the clustered covariance structure, partialing out the fixed effects poses no additional burden after allowing for clustering.

The second contribution of this paper is taking the derived performance bounds for the proposed Lasso variant and using them to provide methods for doing valid inference following variable selection in two canonical models with high dimensional components: the linear instrumental variables (IV) regression with high dimensional instruments and additive fixed effects and the partially linear treatment model with high dimensional controls and additive fixed effects. Inference in these settings is complicated due to the fact that variable selection procedures inevitably make model selection mistakes which may result in invalid inference following model selection; see Pötscher (2009) and Leeb and Pötscher (2008) for examples. It is thus important to offer procedures that are robust to such model selection mistakes. To address this concern, we follow the approach of Belloni, Chen, Chernozhukov, and Hansen (2012) in the IV model and Belloni, Chernozhukov, and Hansen (2014) in the partially linear model making use of the Cluster-Lasso to accommodate within-individual dependence and partialing out of fixed effects. We show that standard inference following these procedures results in inference about model parameters of interest that is uniformly valid within a large class of approximately sparse regression models as long as a clustered covariance estimator is used in estimating the parameters' asymptotic variance. The results of this paper thus allow valid inference about a prespecified, fixed set of model parameters of interest in canonical panel data models with additive fixed effects in the realistic scenario where a researcher is unsure about the exact identities of the relevant set of variables to be included in addition to the variables of interest and the fixed effects.[2]

In addition to theoretical guarantees, we illustrate the performance of the proposed methods through simulation examples and an empirical example. In the simulations, we consider a fixed effects IV model and a conventional linear fixed effects model. The most interesting feature of the simulation results is that the Cluster-Lasso-based procedures perform markedly better than variable selection procedures that do not allow for clustering. This difference in performance suggests that additional modifications of Lasso-type procedures to account for other dependence structures may be worthwhile. In the empirical example, we use our methods to reexamine the Cook and Ludwig (2006) study of the effect of guns on crime using a much broader set of controls than the original paper. Our findings are largely consistent with those of Cook and Ludwig (2006) despite allowing for a much richer set of conditioning information.

The remainder of this paper is organized as follows. In Section 2, we present a variant of the Lasso estimator that is appropriate for high-dimensional panel data models with within group dependence. We present formal results for this Lasso estimator in Section 3. In Section 4, we apply the results from Section 3 to doing inference following variable selection in linear instrumental variables models with fixed effects and the standard linear fixed effects model. Section 5 contains the simulation results, and Section 6 contains the empirical example. We outline a feasible implementation algorithm in the appendix. All proofs of formal results and additional simulation results are available in a further supplementary appendix.

2. Dimension Reduction and Regularization via Lasso Estimation in Panels

An appealing method for estimating sparse high dimensional linear models is the Lasso. Lasso estimates regression coefficients by minimizing a least squares objective plus an ℓ_1 penalty term. We begin with an informal discussion of Lasso in linear models with fixed effects before proceeding with more precise specifications and modeling assumptions. Consider the model

$$y_{it} = x_{it}'\beta + \alpha_i + \epsilon_{it}, \quad i = 1, ..., n, \quad t = 1, ..., T,$$

where y_{it} is an outcome of interest, x_{it} are covariates, α_i are individual specific effects, and ϵ_{it} is an idiosyncratic disturbance term which is mean zero conditional on covariates but may have dependence within an individual. We abstract from issues arising from unbalanced panels for notational convenience but note that the arguments go through immediately provided that the missing observations are missing at random.

[2] In recent complementary work, Ando and Bai (2015a) and Ando and Bai (2015b) consider penalized estimation in panel data models with a rich additive interactive fixed effects structure under fixed model asymptotics.

2.1. Cluster-Lasso Estimation in Panel Models.

The first step in our estimation strategy is to eliminate the fixed effect parameters. For simplicity, we will always consider removing the fixed effects by within individual demeaning but note that removing the fixed effects using other differencing methods could be accommodated using similar arguments. We define

$$\ddot{x}_{it} = x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it}.$$

We define the quantities $\ddot{y}_{it}$ and $\ddot{\epsilon}_{it}$ similarly and note that the double dot notation will signify deviations from within individual means throughout the paper. Eliminating the fixed effects by substracting individual specific means leads to the "within model":

$$\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{\epsilon}_{it}.$$

The Cluster-Lasso coefficient estimate $\widehat{\beta}_L$ is defined by the solution to the following penalized minimization problem on the within model:

$$\widehat{\beta}_L \in \arg\min_b \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\ddot{y}_{it} - \ddot{x}_{it}'b)^2 + \frac{\lambda}{nT}\sum_{j=1}^{p}\widehat{\phi}_j|b_j|. \tag{2.1}$$

Solving the problem (2.1) requires two user-specified tuning parameters: the main penalty level, λ, and covariate specific penalty loadings, $\{\widehat{\phi}_j\}_{j=1}^{p}$. The main penalty parameter dictates the amount of regularization in the Lasso procedure and serves to balance overfitting and bias concerns. The covariate specific penalty loadings $\{\widehat{\phi}_j\}_{j=1}^{p}$ are introduced to allow us to handle data which may be dependent within individual, heteroscedastic, and non-Gaussian. We provide further discussion of the specific choices of penalty parameters in the next subsection.

We will also make use of a post model selection estimator; see for example Belloni and Chernozhukov (2013). The Post-Cluster-Lasso estimator is defined with respect to the variables selected by Cluster-Lasso: $\widehat{I} = \{j : \widehat{\beta}_{Lj} \neq 0\}$. The Post-Cluster-Lasso estimator is simply the least squares estimator subject to the constraint that covariates not selected in the initial Cluster-Lasso regression must have zero coefficients:

$$\widehat{\beta}_{PL} = \underset{b:\ b_j=0\ \forall\ j\notin\widehat{I}}{\text{argmin}} \quad \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\ddot{y}_{it} - \ddot{x}_{it}'b)^2. \tag{2.2}$$

As discussed in Section 3, the selected model $\widehat{I}$ has good properties under regularity conditions and approximate sparsity of the coefficient β. Just as in Belloni and Chernozhukov (2013), good properties of the selected set of variables will then translate into good properties for the Post-Cluster-Lasso estimator.

2.2. **Clustered Penalty Loadings.** An important condition used in proving favorable performance of Cluster-Lasso and inference following Cluster-Lasso-based model selection is the use of penalty loadings and penalty parameters that dominate the score vector in the sense that

$$\frac{\lambda\widehat{\phi}_j}{nT} \geqslant 2c\left|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{\epsilon}_{it}\right| \quad \text{for each } 1 \leqslant j \leqslant p, \tag{2.3}$$

for some constant slack parameter $c > 1$. Belloni, Chernozhukov, and Hansen (2014) refer to condition (2.3) as the "regularization event". Note that the term $\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{\epsilon}_{it}$ intuitively captures the sampling variability in learning about coefficient β_j. The regularization event thus corresponds to selecting penalty parameters large enough to dominate the noise in estimating model coefficients. Looking at the structure of the Lasso optimization problem, (2.1), we can see that (2.3) leads to settings all coefficients whose magnitude is not big enough relative to sampling noise exactly to zero in the Lasso solution. This property makes Lasso-based methods appealing for forecasting and variable selection in sparse models where many of the model parameters can be taken to be zero.

Given the importance of event (2.3) in verifying desirable properties of Lasso-type estimators, it is key that penalty loadings and the penalty level are chosen so that (2.3) occurs with high probability. The intuition for suitable choices can be seen by considering $\widehat{\phi}_j = \phi_j$ where

$$\phi_j^2 = \frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{\epsilon}_{it}\right)^2 = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{x}_{itj}\ddot{x}_{it'j}\ddot{\epsilon}_{it}\ddot{\epsilon}_{it'}.$$

Note that the quantity ϕ_j^2 is a natural measure for the noise in estimating β_j that allows for arbitrary within-individual dependence. With these loadings, we can apply the moderate deviation theorems for self-normalized sums due to Jing, Shao, and Wang (2003) to conclude that

$$\frac{P(\phi_j^{-1}\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{\epsilon}_{it} > m)}{P(N(0,1) > m)} = o(1), \quad \text{uniformly in } |m| = o(n^{1/6}),\ j \in 1, ..., p.$$

It follows from this result and the union bound that setting λ large enough to dominate p standard Gaussian random variables with high-probability, specifically as in (2.6), will implement condition (2.3) using $\widehat{\phi}_j = \phi_j$. This form of loadings is an extension of the loadings considered in Belloni, Chen, Chernozhukov, and Hansen (2012) which apply in settings with non-Gaussian and heteroscedastic but independent data to the present setting where we need to accommodate within-individual dependence.

In practice, the values $\{\phi_j\}_{j=1}^{p}$ are infeasible since they depend on the unobservable $\ddot{\epsilon}_{it}$. To make estimation feasible, we use preliminary estimates of $\ddot{\epsilon}_{it}$, denoted $\widehat{\epsilon}_{it}$, in forming feasible loadings:

$$\widehat{\phi}_j^2 = \frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\ddot{x}_{itj}\widehat{\epsilon}_{it}\right)^2 = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{x}_{itj}\ddot{x}_{it'j}\widehat{\epsilon}_{it}\widehat{\epsilon}_{it'}. \tag{2.4}$$

The $\widehat{\epsilon}_{it}$ can be calculated through an iterative algorithm given in Appendix A which follows the algorithm given in Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2014). We define the Feasible Cluster-Lasso and Feasible Post-Cluster-Lasso estimates as the Cluster-Lasso and Post-Cluster-Lasso estimates using the feasible penalty loadings. A key property of the feasible penalty loadings needed for validity of the approach is that

$$\begin{gathered}\ell\phi_j \leqslant \widehat{\phi}_j \leqslant u\phi_j, \quad \text{with probability } 1-o(1), \\ \text{for some } \ell \to 1 \text{ and } u \leqslant C < \infty, \text{ uniformly for } j=1,...,p.\end{gathered} \tag{2.5}$$

Under this condition and setting

$$\lambda = 2c\sqrt{nT}\Phi^{-1}(1-\gamma/2p) \tag{2.6}$$

with $\gamma = o(1)$, the regularization event (2.3) holds with probability tending to one.

It is worth noting that failure to use the clustered penalty loadings defined in (2.4) (or their infeasible version) can lead to an inflated probability of failure of the regularization event. When this event fails to hold, covariates which are only spuriously related with the outcome have a non-negligable chance of entering the selected model. In the simulation experiments provided in Section 5, we demonstrate how inclusion of such variables can be problematic for post-model-selection inference.

3. Regularity Conditions and Performance Results for Cluster-Lasso Under Grouped Dependence

This section gives conditions under which Cluster-Lasso and Post-Cluster-Lasso attains favorable performance bounds. These bounds are useful in their own right and are important elements in establishing the properties of inference following Lasso variable selection discussed in Section 4. In establishing our formal results, we consider the additive fixed effects model

$$y_{it} = f(w_{it}) + e_i + \epsilon_{it}, \quad \mathrm{E}[\epsilon_{it}|w_{i1},...,w_{iT}] = 0, \quad i=1,...,n, \quad t=1,...,T, \tag{3.7}$$

where e_i represents time invariant individual specific heterogeneity that is allowed to depend on $w_i = \{w_{it}\}_{t=1}^T$ in an unrestricted manner. Throughout, we will assume that $\{y_{it}, w_{it}\}_{t=1}^T$ are i.i.d. across i but do not restrict the within individual dependence. We note that we could allow for data that are i.n.i.d. across i at the cost of complicating the notation and statement of the regularity conditions. Our results will hold under $n \to \infty$, T fixed asymptotics and $n \to \infty$, $T \to \infty$ joint asymptotics.[3]

A key distinction between the analysis in this paper and previous work on Lasso is allowing for within-individual dependence. To aid discussion of this feature, we let

$$\imath_T := T \min_{1 \leqslant j \leqslant p} \frac{\mathrm{E}[\frac{1}{T}\sum_{t=1}^{T} \ddot{x}_{itj}^2 \ddot{\epsilon}_{it}^2]}{\mathrm{E}[\frac{1}{T}(\sum_{t=1}^{T} \ddot{x}_{itj}\ddot{\epsilon}_{it})^2]} = T \min_{1 \leqslant j \leqslant p} \frac{\mathrm{E}[\frac{1}{T}\sum_{t=1}^{T} \ddot{x}_{itj}^2 \ddot{\epsilon}_{it}^2]}{\mathrm{E}[\phi_j^2]}$$

be the *index of information* induced by the "time" or "within-group" dimension. This time information index, $\imath_T$, is inversely related to the strength of within-individual dependence in the scores and can vary between two extreme cases:

- $\imath_T = 1$, no information, corresponding to perfect dependence within the cluster i,
- $\imath_T = T$, maximal information, corresponding to perfect independence within i.

There are many interesting cases between these extremes. A leading case is where $\imath_T \propto T$ which occurs when there is weak dependence within clusters and results in clustering only affecting the constants in the Lasso performance bounds. The case where $\imath_T \propto T^a$ for some $0 \leqslant a < 1$ corresponds to stronger forms of dependence within clusters. Our results will allow for the two extreme cases as well as those falling between these two extremes. It should be noted that our results allow unit-root and other non-stationary behavior of data (x_{it}, ϵ_{it}) across the temporal dimension t in the large n, fixed T setting. However, unit-root and other processes that may result in unbounded unconditional moments are formally not covered in our results under the large n, large T case as we impose boundedness of the first few unconditional moments. We conjecture that under this scenario our methods for choosing penalty parameters still apply and that our proofs can be extended to establish rate and inference results.

We begin the presentation of formal conditions by defining approximately sparse models. Note that the model f as well as the set of covariates w_{it} may depend on the sample size, but we suppress this dependence for notational convenience.

Condition ASM. (Approximately Sparse Model). *The function $f(w_{it})$ is well-approximated by a linear combination of a dictionary of transformations, $x_{it} = X_{nT}(w_{it})$, where x_{it} is a $p \times 1$ vector with $p \gg n$ allowed, and X_{nT} is a measurable map. That is, for each i and t*

$$f(w_{it}) = x_{it}'\beta + r(w_{it}),$$

[3] To accommodate both $n \to \infty$, T fixed asymptotics and $n \to \infty$, $T \to \infty$ joint asymptotics in a simple and unified manner, we maintain strict exogeneity of w_i throughout and do not consider time effects. We note that time effects may be included easily under $n \to \infty$, T fixed asymptotics.

where the coefficient β and the remainder term $r(w_{it})$ satisfy

$$\|\beta\|_0 \leqslant s = o(n\imath_T) \quad \textit{and} \quad \left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} r(w_{it})^2\right]^{1/2} \leqslant A_s = O_{\mathrm{P}}(\sqrt{s/n\imath_T}).$$

We note that the approximation error $r(w_{it})$ is restricted to be of the same order as or smaller than sampling uncertainty in β provided that the true model were known. Because we will mainly be concerned with the within model, we note that it is straightforward to show that $\ddot{y}_{it} = \ddot{f}(w_{it}) + \ddot{\epsilon}_{it}$ satisfies Condition ASM when the original model does.

The next assumption controls the behavior to the empirical Gram matrix. Let $\ddot{M}$ be the $p \times p$ matrix of the sample covariances between the variables $\ddot{x}_{itj}$. Thus,

$$\ddot{M} = \{M_{jk}\}_{j,k=1}^{p}, \quad M_{jk} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{x}_{itk}.$$

In standard regression analysis where the number of covariates is small relative to the sample size, a conventional assumption used in establishing desirable properties of estimators of β is that $\ddot{M}$ has full rank. In the high dimensional setting, $\ddot{M}$ will be singular if $p \geqslant n$ and may have an ill-behaved inverse even when $p < n$. However, good performance of the Lasso estimator only requires good behavior of certain moduli of continuity of $\ddot{M}$. There are multiple formalizations and moduli of continuity that can be considered in establishing the good performance of Lasso; see Bickel, Ritov, and Tsybakov (2009). We focus our analysis on a simple eigenvalue condition that is suitable for most econometric applications. It controls the minimal and maximal m-sparse eigenvalues of $\ddot{M}$ defined as

$$\varphi_{\min}(m)(\ddot{M}) = \min_{\delta\in\Delta(m)} \delta'\ddot{M}\delta \quad \text{and} \quad \varphi_{\max}(m)(\ddot{M}) = \max_{\delta\in\Delta(m)} \delta'\ddot{M}\delta. \tag{3.8}$$

where $\Delta(m) = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leqslant m, \|\delta\|_2 = 1\}$, is the m-sparse subset of a unit sphere.

Condition SE. (Sparse Eigenvalues) *For any $C > 0$, there exist constants $0 < \kappa' < \kappa'' < \infty$, which do not depend on n but may depend on C, such that with probability approaching one, as $n \to \infty$, $\kappa' \leqslant \varphi_{\min}(Cs)(\ddot{M}) \leqslant \varphi_{\max}(Cs)(\ddot{M}) \leqslant \kappa''$.*

Condition SE requires only that certain "small" $Cs \times Cs$ submatrices of the large $p \times p$ empirical Gram matrix are well-behaved. This condition seems reasonable and will be sufficient for the results that follow. Note that we prefer to write the eigenvalue conditions in terms of the demeaned covariates as it is straightforward to show that the conditions continue to hold under data generating processes where the covariates have nonzero within-individual variation. Condition SE could be shown to hold under more primitive conditions by adapting arguments found in Belloni and Chernozhukov (2013) which build upon results in Zhang and Huang (2008) and Rudelson and Vershynin (2008); see also Rudelson and Zhou (2011).

The final condition collects various rate and moment restrictions. The conditions are expressed in terms of demeaned quantities for convenience and will involve the following third moment:

$$\varpi_j = \left(\mathrm{E}\left[\left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \ddot{x}_{itj} \ddot{\epsilon}_{it} \right|^3 \right] \right)^{1/3}.$$

Condition R. (Regularity Conditions) *Assume that for data $\{y_{it}, w_{it}\}$ that are i.i.d. across i, the following conditions hold with x_{it} defined as in Condition ASM with probability $1 - o(1)$:*

(i) $\left(\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}[\ddot{x}_{itj}^2 \ddot{\epsilon}_{it}^2]\right) + \left(\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}[\ddot{x}_{itj}^2 \ddot{\epsilon}_{it}^2]\right)^{-1} = O(1)$,

(ii) $1 \leqslant \max_{1 \leqslant j \leqslant p} \phi_j / \min_{1 \leqslant j \leqslant p} \phi_j = O(1)$,

(iii) $1 \leqslant \max_{1 \leqslant j \leqslant p} \varpi_j / \sqrt{\mathrm{E}\phi_j^2} = O(1)$,

(iv) $\log^3(p) = o(nT)$ and $s \log(p \vee nT) = o(n\imath_T)$,

(v) $\max_{1 \leqslant j \leqslant p} |\phi_j - \sqrt{\mathrm{E}\phi_j^2}| / \sqrt{\mathrm{E}\phi_j^2} = o(1)$.

This condition is sufficient for verifying that Cluster-Lasso and Post-Cluster-Lasso have good model selection and prediction properties under the high-level assumption (2.5) on the availability of valid feasible data loadings. In the appendix we provide additional conditions under which we exhibit validity of data-dependent loadings constructed using an iterative algorithm.

Theorem 1 (Model Selection Properties of Cluster-Lasso and Post-Cluster-Lasso)**.** *Let $\{P_{n,T}\}$ be a sequence of probability laws, such that $\{(y_{it}, w_{it}, x_{it})\}_{t=1}^{T} \sim P_{n,T}$, i.i.d. across i for which n, $T \to \infty$ jointly or $n \to \infty$, T fixed. Suppose that Conditions ASM, SE and R hold for probability measure $\mathrm{P} = \mathrm{P}_{P_{n,T}}$ induced by $P_{n,T}$. Consider a feasible Cluster-Lasso estimator with penalty level (2.6) and loadings obeying (2.5). Then the data-dependent model $\widehat{I}$ selected by a feasible Cluster-Lasso estimator satisfies with probability $1 - o(1)$, $\widehat{s} = |\widehat{I}| \leq Ks$ for some constant $K > 0$ that does not depend on n. In addition, the following relations hold for the Cluster-Lasso estimator ($\widehat{\beta} = \widehat{\beta}_L$) and Post-Cluster-Lasso estimator ($\widehat{\beta} = \widehat{\beta}_{PL}$):*

$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (\ddot{x}_{it}' \widehat{\beta} - \ddot{x}_{it}' \beta)^2 = O_{\mathrm{P}}\left(s \log(p \vee nT) / n\imath_T \right),$$

$$\|\widehat{\beta} - \beta\|_2 = O_{\mathrm{P}}\left(\sqrt{s \log(p \vee nT) / n\imath_T} \right),$$

$$\|\widehat{\beta} - \beta\|_1 = O_{\mathrm{P}}\left(\sqrt{s^2 \log(p \vee nT) / n\imath_T} \right).$$

4. Applications of Cluster-Lasso

The bounds derived in Section 3 allow us to derive the properties of inference methods following variable selection with the Cluster-Lasso. In this section, we use these results to provide two different applications of using Cluster-Lasso to select variables for use in causal inference.

4.1. **Selection of Instruments.**

In this section, we follow Belloni, Chen, Chernozhukov, and Hansen (2012) who consider using Post-Lasso to estimate optimal instruments. Using Lasso-based methods to form first-stage predictions in IV estimation provides a practical approach to obtaining the efficiency gains from using optimal instruments while dampening the problems associated with many instruments. We prove that Cluster-Lasso-based procedures produce first-stage predictions that provide good approximations to the optimal instruments when controlling for individual heterogeneity through fixed effects.

We consider the following model:

$$y_{it} = \alpha d_{it} + e_i + \epsilon_{it} \tag{4.9}$$

$$d_{it} = h(w_{it}) + f_i + u_{it} \tag{4.10}$$

where $\mathrm{E}[\epsilon_{it} u_{it}] \neq 0$ but $\mathrm{E}[\epsilon_{it}|w_{i1}, ..., w_{iT}] = \mathrm{E}[u_{it}|w_{i1}, ..., w_{iT}] = 0$. The extension to d_{it} an $r \times 1$ vector with $r \ll nT$ fixed is straightforward and omitted for convenience. The results also carry over immediately to the case with a small number of included exogenous variables $y_{it} = \alpha d_{it} + x_{it}'\beta + e_i + \epsilon_{it}$ where x_{it} is a $k \times 1$ vector with $k \ll nT$ fixed that will be partialed out with the fixed effects.

We consider estimation of the parameter of interest α, the coefficient on the endogenous regressor, using Cluster-Lasso to select instruments. We assume that the first-stage follows an approximately sparse model with $h(w_{it}) = z_{it}'\pi + r(w_{it})$ where we let $z_{it} = z(w_{it})$ denote a dictionary of transformations of underlying instrument w_{it} and π be a sparse coefficient as in Condition ASM. After eliminating the fixed effect terms through demeaning, the model reduces to

$$\ddot{y}_{it} = \alpha \ddot{d}_{it} + \ddot{\epsilon}_{it} \tag{4.11}$$

$$\ddot{d}_{it} = \ddot{h}(w_{it}) + \ddot{u}_{it} = \ddot{D}_{it} + \ddot{u}_{it} = \ddot{z}_{it}'\pi + \ddot{r}(w_{it}) + \ddot{u}_{it} \tag{4.12}$$

where we set $D_{it} = h(w_{it})$ for notational convenience. By Theorem 1, the Cluster-Lasso estimate of the coefficients on $\ddot{z}_{it}$ when we use $\ddot{z}_{it}$ to predict $\ddot{d}_{it}$, $\widehat{\pi}$, will be sparse with high probability. Letting $\widehat{I}_\pi = \{j : \widehat{\pi}_j \neq 0\}$, the Cluster-Lasso-based estimator of α may be calculated by standard two stage least squares using only the instruments selected by Cluster-Lasso: $\ddot{z}_{it\widehat{I}_\pi} := (\ddot{z}_{itj})_{j\in\widehat{I}_\pi}$. That is, we define the Post-Cluster-Lasso IV estimator for α as

$$\widehat{\alpha} = \widehat{Q}^{-1}\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{D}_{it}\ddot{y}_{it} \text{ where } \widehat{Q} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{d}_{it}\widehat{D}_{it}, \tag{4.13}$$

$\widehat{D_{it}}$ is the fitted value from the regression of $\ddot{d}_{it}$ on $(\ddot{z}_{itj})_{j\in\widehat{I}_\pi}$, and $\widehat{\epsilon}_{it} = \ddot{y}_{it} - \widehat{\alpha}\ddot{d}_{it}$.

We then define an estimator of the asymptotic variance of $\widehat{\alpha}$, which will be used to perform inference for the parameter α after proper rescaling, as

$$\widehat{V} = \widehat{Q}^{-1}\left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{d}_{it}\ddot{d}_{it'}\widehat{\epsilon}_{it}\widehat{\epsilon}_{it'}\right)\widehat{Q}^{-1}. \tag{4.14}$$

Scaled appropriately, the estimate $\widehat{V}$ will be close to the quantity

$$V = \frac{\imath_T^D}{T}Q^{-1}\Omega Q^{-1}, \quad \text{with probability } 1 - o(1)$$

where

$$Q = \mathrm{E}[\frac{1}{T}\sum_{t=1}^{T}\ddot{D}_{it}^2], \quad \Omega = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathrm{E}[\ddot{D}_{it}\ddot{D}_{it'}\ddot{\epsilon}_{it}\ddot{\epsilon}_{it'}].$$

Finally, it is convenient to define the following quantities that are useful in discussing formal conditions for our estimation procedure. We define appropriate moments and information indices analogous to those used to derive properties of Cluster-Lasso and Post-Cluster-Lasso. For any arbitrary random variables, $A = \{A_{it}\}_{i\leqslant n, t\leqslant T}$, define

$$\phi^2(A) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_{it}\right)^2, \quad \varpi(A) = \mathrm{E}\left[\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_{it}\right|^3\right]^{1/3}, \quad \imath_T(A) = T\frac{\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T}A_{it}^2\right]}{\mathrm{E}\left[\phi^2(A)\right]}.$$

For use in the instrumental variables estimation, we let

$$\begin{aligned}
\phi_j^2 &= \phi^2(\{\ddot{z}_{itj}\ddot{u}_{it}\}), \quad \varpi_j = \varpi(\{\ddot{z}_{itj}\ddot{u}_{it}\}), \qquad \imath_T = \min_{j\leqslant p}\imath_T(\{\ddot{z}_{itj}\ddot{u}_{it}\})\\
\phi_D^2 &= \phi^2(\{\ddot{D}_{it}\ddot{\epsilon}_{it}\}), \quad \varpi_D = \varpi(\{\ddot{D}_{it}\ddot{\epsilon}_{it}\}), \qquad \imath_T^D = \imath_T(\{\ddot{D}_{it}\ddot{\epsilon}_{it}\})\\
\phi_{z_jd}^2 &= \phi^2(\{\ddot{z}_{itj}\ddot{d}_{it}\}), \quad \varpi_{z_jd} = \varpi(\{\ddot{z}_{itj}\ddot{d}_{it}\}), \quad \imath_T^{z_jd} = \imath_T(\{\ddot{z}_{itj}\ddot{d}_{it}\})\\
\phi_{z_j\epsilon}^2 &= \phi^2(\{\ddot{z}_{itj}\ddot{\epsilon}_{it}\}), \quad \varpi_{z_j\epsilon} = \varpi(\{\ddot{z}_{itj}\ddot{\epsilon}_{it}\}), \quad \imath_T^{z_j\epsilon} = \imath_T(\{\ddot{z}_{itj}\ddot{\epsilon}_{it}\})
\end{aligned}$$

To derive asymptotic properties of these estimators, we will make use of the following condition in addition to those assumed in Section 3.

Condition SMIV

(i) Sufficient conditions for Post-Cluster-Lasso: ASM, SE, R hold for model 4.10.

(ii) Sufficient conditions for asymptotic normality of $\widehat{\alpha}$ and consistency of $\frac{\imath_T^D}{T}\widehat{V}$:

(a) $\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T \ddot{D}_{it}^2\right]$, $\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T \ddot{\epsilon}_{it}^2\ddot{D}_{it}^2\right]$, $\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^T \ddot{d}_{it}^2\right)^2\right]$ *are bounded uniformly from above and away from zero, uniformly in* n, T. *Additionally,* $\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^T \ddot{\epsilon}_{it}^2\right)^q\right] = O(1)$ *for some* $q > 4$.

(b) $\varpi_D/\sqrt{\mathrm{E}\phi_D^2} = O(1)$, $\max_{1\leqslant j\leqslant p} \varpi_{z_j\epsilon}/\sqrt{\mathrm{E}\phi_{z_j\epsilon}^2} = O(1)$,

(c) $\max_j \frac{\imath_T^{z_j\epsilon}}{T}\phi_{z_j\epsilon}^2 = O_P(1)$, $\frac{1}{T}\phi_{dD}^2 = O_{\mathrm{P}}(1)$, $\max_j \frac{1}{T}\phi_{z_jd}^2 = O_P(1)$

(d) $\frac{s^2\log^2(p\vee nT)}{n\imath_T}\max\{1, \max_{1\leqslant j\leqslant p}\frac{\imath_T^D}{\imath_T^{z_j\epsilon}}\} = o(1)$ *and* $\frac{\imath_T^D}{\imath_T}n^{2/q}\frac{s\log(p\vee nT)}{n} = o(1)$

The conditions assumed in Condition SMIV are fairly standard. Outside of moment conditions, the main restriction in Condition SMIV is condition (ii)(a) that guarantees that the parameter α would be strongly identified if $\ddot{D}_{it}$ could be observed. Coupled with the approximately sparse model, this condition implies that using a small number of the variables in z_{it} is sufficient to strongly identify α which rules out the case of weak-instruments as in Staiger and Stock (1997) and many-weak-instruments as in Chao, Swanson, Hausman, Newey, and Woutersen (2012).[4]

Theorem 2 (Estimation and Inference in IV Models)**.** *Uniformly over all sequences* $\{\mathrm{P}_{n,T}\}$ *for which* $\{(y_{it}, x_{it}, z_{it})\}_{t=1}^T \sim \mathrm{P}_{n,T}$*, i.i.d. across i, for which the instrumental variable model holds, and for which condition SMIV holds,*[5]*,*

$$\sqrt{n\imath_T^D}V^{-1/2}(\widehat{\alpha} - \alpha) \xrightarrow{d} N(0,1) \quad \text{and} \quad V - \frac{\imath_T^D}{T}\widehat{V} \xrightarrow{P} 0.$$

This theorem verifies that the IV estimator formed with instruments selected by Cluster-Lasso in a linear IV model with fixed effects is consistent and asymptotically normal. In addition, one can use the result with $\widehat{V}$ defined in (4.14), which is simply the usual clustered standard error estimator (Arellano, 1987), to perform valid inference for α following instrument selection. Note that this inference will be valid uniformly over a large class of data generating processes which includes cases where perfect instrument selection is impossible.

4.2. **Selection of Control Variables.** A second strategy for identifying structural effects in economic research is based on assuming that variables of interest are as good as randomly assigned conditional on time varying observables and time invariant fixed effects. Since this approach relies on including the right set of time varying observables, a practical problem

[4] See also Hansen and Kozbur (2014) who consider many-weak-instruments in a $p > n$ setting.

[5] More precisely, the convergence holds uniformly over sequences satisfying Condition SMIV, with the same implied constants with $n,\ T \to \infty$ jointly or $n \to \infty,\ T$ fixed.

researchers face is the choice of which control variables to include in the model. The high dimensional framework provides a convenient setting for exploring data-dependent selection of control variables. In this section, we consider the problem of selecting a set of variables to include in a linear model from a large set of possible control variables in the presence of unrestricted individual specific heterogeneity.

The structure of the Lasso optimization problem ensures that any estimated coefficient that is not set to zero can be reliably differentiated from zero relative to estimation noise when (2.3) holds while any coefficient that can not be distinguished reliably from zero will be estimated to be exactly zero. This property complicates inference after model selection in approximately sparse models which may have a set of variables with small but non-zero coefficients in addition to strong predictors. In this case, satisfaction of condition (2.3) will result in excluding variables with small but non-zero coefficients which may lead to non-negligible omitted variables bias and irregular sampling behavior of estimates of parameters of interest. This intuition is formally developed in Pötscher (2009) and Leeb and Pötscher (2008). Offering solutions to this problem with fully independent data is the focus of a number of recent papers; see, for example, Belloni, Chernozhukov, and Hansen (2010); Belloni, Chen, Chernozhukov, and Hansen (2012); Zhang and Zhang (2014); Belloni, Chernozhukov, and Hansen (2013); Belloni, Chernozhukov, and Hansen (2014); van de Geer, Bühlmann, Ritov, and Dezeure (2014); Javanmard and Montanari (2014); Farrell (2013); and Belloni, Chernozhukov, Fernández-Val, and Hansen (2014). In this section, we focus on extending the approach of Belloni, Chernozhukov, and Hansen (2014) to the panel setting with dependence within individuals.

To be precise, we consider estimation of the parameter α in the partially linear additive fixed effects panel model:

$$y_{it} = d_{it}\alpha + g(z_{it}) + e_i + \zeta_{it}, \qquad \mathrm{E}[\zeta_{it} \mid z_{i1}, ..., z_{iT}, d_{i1}, ..., d_{iT}, e_i] = 0, \tag{4.15}$$

$$d_{it} = m(z_{it}) + f_i + u_{it}, \qquad \mathrm{E}[u_{it} \mid z_{i1}, ..., z_{iT}, f_i] = 0, \tag{4.16}$$

where y_{it} is the outcome variable, d_{it} is the policy/treatment variable whose impact α we would like to infer. The analysis extends easily to the case where d_{it} is an $r \times 1$ vector where r is fixed and is omitted for convenience. z_{it} represents confounding factors on which we need to condition, e_i and f_i are fixed effects which are invariant across time, and ζ_{it} and u_{it} are disturbances that are independent of each other. Data are assumed independent across i, and dependence over time within individual is largely unrestricted.

The confounding factors z_{it} affect the policy variable via the function $m(z_{it})$ and the outcome variable via the function $g(z_{it})$. Both of these functions are unknown and potentially complicated. We use linear combinations of control terms $x_{it} = P(z_{it})$ to approximate

$g(z_{it}) = x_{it}'\beta_g + r_g(z_it)$ and $m(z_{it}) = x_{it}'\beta_m + r_m(z_it)$ with $r_g(z_{it})$ and $r_m(z_{it})$ being approximation errors. In order to allow for a flexible specification and incorporation of pertinent confounding factors, we allow the dimension, p, of the vector of controls, $x_{it} = P(z_{it})$, to be large relative to the sample size. Upon substituting these approximation into (4.15) and (4.16) and removing fixed effects, we are essentially left with a conventional linear fixed effects model with a high dimensional set of potential confounding variables:

$$\ddot{y}_{it} = \ddot{d}_{it}\alpha + \ddot{x}_{it}'\beta_g + \ddot{r}_g(z_{it}) + \ddot{\zeta}_{it}, \tag{4.17}$$

$$\ddot{d}_{it} = \ddot{x}_{it}'\beta_m + \ddot{r}_m(z_{it}) + \ddot{u}_{it}. \tag{4.18}$$

Informative inference about α is not possible in this model without imposing further structure since we allow for $p > n$ elements in x_{it}. The additional structure is added by assuming that condition ASM applies to both $g(z_{it})$ and $m(z_{it})$ which implies that exogeneity of d_{it} may be taken as given once one controls linearly for a relatively small number, $s < n$, of the variables in x_{it} whose identities are *a priori* unknown. Under this condition, estimation of α may then proceed by using variable selection methods to choose a set of relevant control variables from among the set $\ddot{x}_{it}$ to use in estimating (4.17).

To estimate α in this environment, we adopt the post-double-selection method of Belloni, Chernozhukov, and Hansen (2014). This method proceeds by first substituting (4.18) into (4.17) to obtain predictive relationships for the outcome $\ddot{y}_{it}$ and the treatment $\ddot{d}_{it}$ in terms of only control variables:

$$\ddot{y}_{it} = \ddot{x}_{it}'\pi + \ddot{r}_{RF}(z_{it}) + \ddot{v}_{it}, \tag{4.19}$$

$$\ddot{d}_{it} = \ddot{x}_{it}'\beta_m + \ddot{r}_m(z_{it}) + \ddot{u}_{it}. \tag{4.20}$$

We then use two variable selection steps. Cluster-Lasso is applied to equation (4.19) to select a set of variables that are useful for predicting $\ddot{y}_{it}$; we collect the controls x_{itj} for which $\widehat{\pi}_j \neq 0$ in the set $\widehat{I}_{RF}$. Cluster-Lasso is then applied to equation (4.20) to select a set of variables that are useful for predicting $\ddot{d}_{it}$; we again collect the controls x_{itj} for which $\widehat{\beta}_{m,j} \neq 0$ in the set $\widehat{I}_{FS}$. The set of controls that will be used is then defined by the union $\widehat{I} = \widehat{I}_{FS} \cup \widehat{I}_{RF}$. Estimation and inference for α may then proceed by ordinary least squares estimation of $\ddot{y}_{it}$ on $\ddot{d}_{it}$ and the set of controls in $\widehat{I}$ using conventional clustered standard errors (Arellano, 1987).

We present additional moment and rate conditions before stating a result which can be used for performing inference about α. We again define several moments using the same notation introduced before condition SMIV:

$$\phi^2_{x_ju} = \phi^2_{j,FS} = \phi^2(\{\ddot{x}_{itj}\ddot{u}_{it}\}), \varpi_{x_ju} = \varpi_{j,FS} = \varpi(\{\ddot{x}_{itj}\ddot{u}_{it}\}), \imath_T^{x_ju} = \imath_T(\{\ddot{x}_{itj}\ddot{u}_{it}\}), \imath_T^{FS} = \min_{j\leqslant p} \imath_T^{x_ju}$$

$$\phi^2_{x_jv} = \phi^2_{j,RF} = \phi^2(\{\ddot{x}_{itj}\ddot{v}_{it}\}), \varpi_{x_jv} = \varpi_{j,RF} = \varpi(\{\ddot{x}_{itj}\ddot{v}_{it}\}), \imath_T^{x_jv} = \imath_T(\{\ddot{x}_{itj}\ddot{v}_{it}\}), \imath_T^{RF} = \min_{j\leqslant p} \imath_T^{x_jv}$$

$$\phi^2_{u\zeta} = \phi^2(\{\ddot u_{it}\ddot\zeta_{it}\}),\ \varpi_{u\zeta} = \varpi(\{\ddot u_{it}\ddot\zeta_{it}\}),\ \imath_T^{u\zeta} = \imath_T(\{\ddot u_{it}\ddot\zeta_{it}\})$$

$$\phi^2_{x_j\zeta} = \phi^2(\{\ddot x_{itj}\ddot\zeta_{it}\}),\ \varpi_{x_j\zeta} = \varpi(\{\ddot x_{itj}\ddot\zeta_{it}\}),\ \imath_T^{x_j\zeta} = \imath_T(\{\ddot x_{itj}\ddot\zeta_{it}\})$$

$$\phi^2_{ud} = \phi^2(\{\ddot u_{it}\ddot d_{it}\}),\ \varpi_{ud} = \varpi(\{\ddot u_{it}\ddot d_{it}\}),\ \imath_T^{ud} = \imath_T(\{\ddot u_{it}\ddot d_{it}\})$$

Condition SMPLM

(i) Sufficient conditions for Post-Cluster-Lasso: ASM, SE, R hold for models 4.19 and 4.20.

(ii) Sufficient conditions for asymptotic normality of $\widehat\alpha$ and consistency of $\frac{\imath_T^D}{T}\widehat V$:

(a) $Q = \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T \ddot u_{it}^2\right]$, $\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T \ddot u_{it}^2\ddot\zeta_{it}^2\right]$, $\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^T \ddot u_{it}^2\right)^2\right]$ *are bounded uniformly from above and away from zero, uniformly in* n, T. *Additionally,* $\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^T \ddot\zeta_{it}^2\right)^q\right] = O(1)$, $\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^T \ddot u_{it}^2\right)^q\right] = O(1)$ and $\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^T \ddot d_{it}^2\right)^q\right] = O(1)$ for some $q > 4$. $|\alpha| \leqslant B < \infty$.

(b) $\varpi_{u\zeta}/\sqrt{\mathrm{E}\phi^2_{u\zeta}} = O(1)$, $\max\limits_{1\leqslant j\leqslant p} \varpi_{x_j\zeta}/\sqrt{\mathrm{E}\phi^2_{x_j\zeta}} = O(1)$, $\max\limits_{1\leqslant j\leqslant p} \varpi_{x_ju}/\sqrt{\mathrm{E}\phi^2_{x_ju}} = O(1)$.

(c) $\max_j \frac{\imath_T^{x_j\zeta}}{T}\phi^2_{x_j\zeta} = O_P(1)$, $\frac{\imath_T^{u\zeta}}{T}\phi^2_{u\zeta} = O_{\mathrm{P}}(1)$, $\max_j \frac{\imath_T^{x_ju}}{T}\phi^2_{x_ju} = O_P(1)$, $\frac{1}{T}\phi^2_{ud} = O_{\mathrm{P}}(1)$.

(d) $\frac{\imath_T^{u\zeta}}{\min\{\imath_T^{RF},\imath_T^{FS},\min_j\{\imath_T^{x_j\zeta}\}\}}\left(s + n^{2/q}\right)\left(\max_{i,t,j}\ddot x_{itj}^2\right)\frac{s\log^2(p\vee nT)}{n} = o_{\mathrm{P}}(1)$.

Finally, we define the following variance estimators for the post double selection procedure:

$$\widehat V_n = \widehat Q^{-1}\widehat\Omega\widehat Q^{-1} \text{ with } \widehat Q = \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T \widehat u_{it}^2, \quad \widehat\Omega = \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\sum_{t'=1}^T \widehat u_{it}\widehat u_{it'}\widehat\zeta_{it}\widehat\zeta_{it'},$$

and $\widehat u_{it} = \ddot d_{it} - \ddot x_{it}'\widehat\beta_m$, $\widehat\zeta_{it} = \ddot y_{it} - \widehat\alpha\ddot d_{it} - \ddot x_{it}'\widehat\beta_g$, $\widehat\beta_m = \underset{b:\ b_j=0\ \forall\ j\notin\widehat I}{\operatorname{argmin}} \sum_{i=1}^n\sum_{t=1}^T(\ddot d_{it} - \ddot x_{it}'b)^2$, $(\widehat\alpha, \widehat\beta_g')' = \underset{(a,b):\ b_j=0\ \forall\ j\notin\widehat I}{\operatorname{argmin}} \sum_{i=1}^n\sum_{t=1}^T(\ddot y_{it} - a\ddot d_{it} - \ddot x_{it}'b)^2$.

Theorem 3 (Estimation and Inference on Treatment Effects)**.** *Uniformly over all sequences* $\{\mathrm{P}_n\}$ *for which* $\{(y_{it}, d_{it}, x_{it})\}_{t=1}^T \sim \mathrm{P}_n$, *i.i.d. across* i, *for which Condition SMPLM holds,*[6] *the Post-Double-Cluster-Lasso estimator* $\widehat\alpha$ *satisfies*

$$\sqrt{n\imath_T^{u\zeta}}V^{-1/2}(\widehat\alpha - \alpha) \xrightarrow{d} N(0,1) \text{ and } V - \frac{\imath_T^{u\zeta}}{T}\widehat V \xrightarrow{\mathrm{P}} 0.$$

This theorem verifies that the OLS estimator which regresses $\ddot y_{it}$ on $\ddot d_{it}$ and the union of variables selected by Cluster-Lasso from (4.19) and (4.20) is consistent and asymptotically normal with asymptotic variance that can be estimated with the conventional clustered standard

[6] More precisely, the convergence holds uniformly over sequences satisfying Condition SMPLM, with the same implied constants with $n,\ T \to \infty$ jointly or $n \to \infty,\ T$ fixed.

error estimator. Inference based on this result will be valid uniformly over a large class of data generating processes which includes cases where perfect variable selection is impossible.

5. Simulation Examples

The results in the previous sections suggest that Cluster-Lasso based estimates should have good estimation and inference properties in panel models with individual specific heterogeneity provided the sample size n is large. In this section, we provide simulation evidence about the performance of our asymptotic approximation for inference about structural parameters in IV models with fixed effects and many instruments and linear fixed effects models when Cluster-Lasso is used for variable selection.

5.1. **Simulation 1: IV.** The first simulation illustrates the performance of the Cluster-Lasso based IV estimator in a simple instrumental variables model with fixed effects and many instruments. In our simulation experiments, we generate data from the linear IV model

$$\begin{aligned} y_{it} &= \alpha d_{it} + e_i + \epsilon_{it} \\ d_{it} &= z_{it}'\pi + f_i + u_{it}. \end{aligned}$$

We generate disturbances according to

$$\begin{aligned} \epsilon_{it} &= \rho_\epsilon \epsilon_{it-1} + \nu_{1,it} \\ u_{it} &= \rho_u u_{it-1} + \nu_{2,it} \end{aligned} \quad \text{where} \quad \begin{pmatrix} \nu_{1,it} \\ \nu_{2,it} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_\nu \\ \rho_\nu & 1 \end{pmatrix} \right) \quad \text{iid}$$

with initial conditions for ϵ_{it} and u_{it} drawn from their stationary distribution. We generate the individual heterogeneity e_i for $i = 1, ..., n$ as correlated normal random variables with $\mathrm{E}[e_i] = 0$, $\mathrm{Var}(e_i) = \frac{4}{T}$, and $\mathrm{Corr}(e_i, e_j) = .5^{|i-j|}$ for all i and j. We set $f_i = e_i$. We draw the instruments conditional on the fixed effects from

$$z_{itj} = e_i + \rho_z z_{i(t-1)j} + \varphi_{itj} \quad \text{for} \quad t > 1 \quad \text{and} \quad z_{i1j} = \frac{e_i}{1-\rho_z} + \sqrt{\frac{1}{1-\rho_z^2}}\varphi_{i1j}$$

where φ_{itj} are normal random variables with $\mathrm{E}[\varphi_{itj}] = 0$, $\mathrm{Var}(\varphi_{itj}) = 1$, and $\mathrm{Corr}(\varphi_{itj}, \varphi_{itk}) = .5^{|j-k|}$ that are independent across i and t. In all of our simulations, we set $\rho_\epsilon = \rho_u = \rho_z = .8$, and we set $\rho_\nu = .5$. We also set $\alpha = .5$. We redraw the disturbances ϵ and u at each simulation replication but condition on one realization of the fixed effects and instruments. We consider different sample sizes set to $n = 50, 100, 150, 200$ all with $T = 10$. Note that the instruments are not valid without conditioning on the fixed effects within this structure. The fixed effects are also dense in the sense that most of the generated effects will be small but non-zero.

The final features of the design are the number of instruments and the structure of the coefficients on the instruments, π. We define the coefficient vector π as

$$\pi_j = (-1)^{j-1}\left(\frac{1}{\sqrt{s}}1_{\{j\leqslant s\}} + \frac{1}{j^2}1_{\{j>s\}}\right), \quad s = \lfloor \frac{1}{2} n^{1/3} \rfloor$$

for $1 \leqslant j \leqslant p$ where $\lfloor a \rfloor$ returns the integer part of a.[7] This design is not exactly sparse due to the presence of the variables with coefficients $\frac{1}{j^2}$ but is approximately sparse as required by condition ASM. Finally, we consider two different numbers of instruments, $p = n \times (T-2)$ and $p = n \times (T+2)$, for each sample size and design of first-stage coefficients.

We report results from five different estimators. We consider IV estimates based on variables selected using the clustered penalty loadings developed in this paper (Clustered Loadings). As a comparison, we also consider IV estimates based on variables selected using the loadings that are valid with heteroscedastic and independent data from Belloni, Chen, Chernozhukov, and Hansen (2012) (Heteroscedastic Loadings). In cases with $p < nT$, we report estimates using 2SLS on the full set of instruments (All). Finally, we consider two different infeasible oracle estimators. The first oracle knows the value of the coefficients π (Oracle) while the second also knows the exact values of the fixed effects (FE Oracle). Thus, both oracle estimators use a single instrument that uses the true values of the first stage coefficients, $z_{it}'\pi$. The difference between the two is that the fixed effects are removed by taking differences of all variables from within-individual means in the Oracle results while the true values of the FE are directly subtracted from y_{it} and d_{it} in the FE Oracle results.

The results are based on 1000 simulations for each setting described above. For results based on All, Heteroscedastic Loadings, Clustered Loadings, and Oracle, the fixed effects are treated as unknown parameters and eliminated by taking deviations from within-individual means. For each estimator, we report mean bias, root mean squared error, and rejection rates for a 5%-level test of $H_0 : \alpha = .5$ using both clustered standard errors and heteroscedastic standard errors.[8] In some of the simulation replications, the IV estimator using variables selected by Lasso is undefined as Lasso sets all coefficients to zero. In such a case, we record a failure to reject the null which is a conservative alternative to applying the Sup-score statistic described in Belloni, Chen, Chernozhukov, and Hansen (2012). Mean bias and root-mean-square-error for Lasso-based estimates are calculated conditional on Lasso selecting at least one instrument.

[7] In the supplementary appendix, we report further simulation results based on two additional structures for the coefficients on the instruments. In these simulations, our procedure continues to perform well even in somewhat adversarial conditions.

[8] Since moments of IV estimators may not exist, we calculate truncated bias and truncated RMSE, truncating at $\pm 10,000$.

The results for estimation of α are reported in Table 1. The two oracle estimators provide infeasible benchmarks. Looking at these results, we see that IV based on the infeasible instruments formed using the true values of the first-stage coefficients perform well in the designs considered. As expected given the well-known properties of 2SLS, the 2SLS estimates using the full set of instruments when $p < nT$ exhibit large bias relative to standard error, large RMSE, and produce tests that suffer from large size distortions.

The Lasso-based results where we do variable selection using loadings that are appropriate under independence but ignore within-individual dependence are quite interesting. This approach performs relatively well compared to naive 2SLS using all of the instruments. However, using instruments selected by Lasso with loadings that ignore the dependence produces an IV estimator of α that has a substantial bias and results in tests that have large size distortions even when clustered standard errors are applied. The presence of this bias illustrates the point that care must be taken when selecting instruments for a post model selection analysis. In general, $\mathrm{E}[z_{itj}\epsilon_{it}| \quad j \quad \text{selected}] \neq 0$ though the difference from zero is ignorable when (2.3) occurs. However, in the absence of the regularization event (2.3), this conditional expectation can be large which introduces a type of "endogeneity" bias as the selected instruments are effectively invalid. We see this behavior when using the heteroscedastic loadings in the designs we consider as these loadings produce smaller penalty levels than the appropriate clustered loadings which results in the spurious inclusion of instruments.

Finally, we see that IV based on instruments selected by Cluster-Lasso clearly dominates the other feasible procedures in the simulation designs considered. Using this procedure produces tests that have approximately correct size that is comparable to size of tests based on both oracle models considered. We also see that the performance for Bias, RMSE, and size of tests is similar to the infeasible Oracle benchmark. Overall, these results are favorable in that using Cluster-Lasso to select instruments outperforms the other methods explored here.

5.2. **Simulation 2: Linear Model.** In this simulation, we consider estimation of a coefficient on a variable of interest in a standard linear fixed effects model. Specifically, we generate data according to the model

$$\begin{aligned} y_{it} &= \alpha d_{it} + z_{it}'\beta + e_i + \epsilon_{it} \\ d_{it} &= z_{it}'\gamma + f_i + u_{it}. \end{aligned}$$

We generate disturbances according to

$$\begin{aligned} \epsilon_{it} &= \rho_\epsilon \epsilon_{it-1} + \nu_{1,it} \\ u_{it} &= \rho_u u_{it-1} + \nu_{2,it} \end{aligned} \quad \text{where} \quad \begin{pmatrix} \nu_{1,it} \\ \nu_{2,it} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \quad \text{iid}$$

with initial conditions for ϵ_{it} and u_{it} drawn from their stationary distribution. We generate e_i, f_i, and z_{it} exactly as in Section 5.1, so we omit the details for brevity. We again set

$\rho_\epsilon = \rho_u = .8$ and set $\alpha = .5$. We redraw the disturbances ϵ and u at each simulation replication but condition on one realization of the fixed effects and controls. We use sample sizes set to $n = 50, 100, 150, 200$ with $T = 10$. As in Section 5.1, we we specify the coefficient vectors β and γ as

$$\gamma_j = \beta_j = (-1)^{j-1}\left(\frac{1}{\sqrt{s}}1_{\{j\leqslant s\}} + \frac{1}{j^2}1_{\{j>2\}}\right), \quad s = \lfloor\frac{1}{2}n^{1/3}\rfloor$$

for $1 \leqslant j \leqslant p$ where $\lfloor a\rfloor$ returns the integer part of a.[9] Again, we consider $p = n \times (T-2)$ and $p = n \times (T+2)$.

For this simulation, we consider six estimators of α. When $p \leqslant nT$, we use the conventional fixed effects estimator including all the variables in z_{it} (All). We use the post-double-selection method with penalty loadings appropriate for independent, heteroscedastic data in each Lasso stage (Heteroscedastic Loadings) and with our clustered loadings in each Lasso stage (Clustered Loadings). We also consider a post-double-selection estimator which includes the fixed effects in the set of variables over which selection occurs (Select over FE) using the approach of Kock (2014). We also consider two oracle estimators. The first oracle knows the values of the coefficients β and γ (Oracle) while the second also knows the exact values of the fixed effects (FE Oracle). The Oracle estimate of α is thus obtained by regressing $\ddot{y}_{it} - \ddot{z}_{it}'\beta$ onto $\ddot{d}_{it} - \ddot{z}_{it}'\gamma$ while the FE Oracle estimate of α is obtained by regressing $y_{it} - z_{it}'\beta - e_i$ onto $d_{it} - z_{it}'\gamma - f_i$. As before, the results are based on 1000 simulation replications for each setting; and we report mean bias, root mean squared error, and rejection rates for a 5%-level test of $H_0 : \alpha = .5$ using both clustered standard errors and heteroscedastic standard errors for each estimator.

The results for the partially linear model simulations are reported in Table 2. The two oracle estimators provide infeasible benchmarks and unsurprisingly produce estimators with small bias and RMSE and tests with reasonable size as long as clustered standard errors are used as is conventional in the literature, e.g. Bertrand, Duflo, and Mullainathan (2004). In all simulations, estimates using the full set of controls when feasible have small bias but large variability leading to large RMSE relative to oracle estimates. Tests based on estimators using the full set of controls are also badly size distorted regardless of whether heteroscedastic or clustered standard errors are used. This distortion results from the difficulty in robustly estimating standard errors when many variables are included; see, e.g., Cattaneo, Jansson, and Newey (2010). This feature suggests that one may not wish to simply include many controls without regularization even when possible.

Estimates based on the double selection method using Lasso with penalty loading appropriate under heteroscedasticity and independence or using Lasso to also select over the fixed

[9] In the supplementary appendix, we report further simulation results based on two additional structures for the coefficients β and γ. In these simulations, our procedure continues to perform well even in somewhat adversarial conditions.

effects perform better than simply including all controls in the $p < nT$ case but tend to perform poorly in terms of bias and coverage probabilities. The bias and poor coverage properties of the estimator that attempts to select over the fixed effects is due to the difficulties in performing selection over the dense part of the model and shows the importance of eliminating fixed effect parameters via demeaning or differencing. These difficulties arise because sparsity provides a poor approximation to the true fixed effects structure. Note that a dense model over unobserved heterogeneity where heterogeneity matters differentially for each individual seems quite reasonable in many economic applications and suggests that attempting to select over fixed effects may result in undesirable features at least when inference about model parameters is the goal of the empirical analysis.

We find it more surprising that using heteroscedastic penalty loadings also leads to noticeable bias and a distortion in statistical size. The heteroscedastic loadings lead to less penalization in our designs which result in inclusion of a few spurious variables. Usual intuition for linear models suggests that including a few extra variables has little impact on say ordinary least squares estimates of parameters of interest. The difficulty arises because the spuriously included variables are not included at random but are exactly those variables with little to no impact that are most highly correlated to the noise and are not properly screened out because the penalty is too low for (2.3) to be a reliable guide. Choosing the variables most highly correlated to the noise then yields that $\mathrm{E}[x_{itj}\epsilon_{it}| \quad j \quad \text{selected}]$ is not negligible due to the use of incorrect penalty loadings leading to biased estimation just as in the instrumental variables case.

Finally, we again see that basing estimation and inference for α on the post-double-selection method using clustered penalty loadings clearly dominates the other feasible procedures in the simulation designs considered. This procedure yields an estimator with RMSE comparable to the oracles across all designs considered. We also see that feasible inference based on this procedure does a relatively good job controlling size across all designs considered. Overall, these results are favorable to Lasso-based variable selection using clustered penalty loadings after partialing out fixed effects and suggests that these methods may offer useful tools to empirical researchers faced with high dimensional panel data.

6. Empirical Example: The Social Cost of Gun Ownership

In the earlier sections, we provided results on the performance of Lasso as a model selection device for panel data models with fixed effects and discussed how to apply Lasso to problems of economic interest in such settings. In this section, we demonstrate the use of Cluster-Lasso by reexamining the Cook and Ludwig (2006) study of the impact of gun ownership on crime.

We briefly review Cook and Ludwig (2006) before presenting the results using the methods described in this paper.

Cook and Ludwig (2006) give several arguments suggesting that gun ownership levels may impose externalities on a community. On the one hand, widespread prevelance of guns can act a deterrent to criminal activity. On the other hand, higher gun prevelance in the general population can lead to higher gun ownership among dangerous people, perhaps through theft or illegal sales, which may lead to an increase in crime. Thus, it is unclear whether the net effect of guns is positive or negative. To investigate the impact of guns, Cook and Ludwig (2006) estimate the effect of gun prevelance on several measures of crime rates. In this example, we revisit their estimation of the effect of gun prevelance on homicide rates.

A major contribution of Cook and Ludwig (2006) is to provide an improved measure of gun ownership in order to get more accurate estimates of the social costs of gun prevalence. Previously, several authors had obtained conflicting estimates for the effect of interest; see, e.g., Lott (2000) and Duggan (2001). Because exact gun-ownership numbers in the U.S. are difficult to obtain, Cook and Ludwig (2006) instead use the fraction of suicides committed with a firearm (abbreviated FSS) within a county as a proxy for county-level gun ownership rates. Cook and Ludwig (2006) argue that if guns are prevalent within a county, then they should be more accessible for the purpose of suicide. They show that their proxy for gun prevelance, FSS, matches up with survey data directly measuring gun ownership from the General Social Survey better than previously used measures. In our analysis, we take it as given that FSS provides a useful measure of gun ownership and that learning the causal effect of FSS is an interesting goal. We thus abstract from any further measurement issues in order to give a clear illustration of our methods.

The main strategy employed by Cook and Ludwig (2006) to estimate causal effects of gun prevalence is to exploit differences in gun ownership across counties and over time. Cook and Ludwig (2006) construct a panel of 195 large United States counties between the years 1980 through 1999 and use this data to estimate linear fixed effects models of the form

$$\log Y_{it} = \beta_0 + \beta_1 \log \text{FSS}_{it-1} + X_{it}'\beta_X + \alpha_i + \delta_t + \epsilon_{it} \tag{6.21}$$

where α_i and δ_t are respectively unobserved county and year effects that will be treated as parameters to be estimated, X_{it} are additional covariates meant to control for any factors related to both gun ownership rates and crime rates that vary across counties and over time, and Y_{it} is one of three dependent variables: the overall homicide rate within county i in year t, the firearm homicide rate within county i in year t, or the non-firearm homicide rate within county i in year t. Cook and Ludwig (2006) consider controls, X_{it}, for percent African American, percent of households with female head, nonviolent crime rates, and percent of the population that lived in the same house five years earlier.

Interpreting the estimated effect of gun prevelance as measured by FSS as causal relies on the belief that there are no variables associated both to crime rates and FSS that are not included in (6.21). The inclusion of county and time fixed effects accounts for any aggregate macroeconomic conditions that affect all counties uniformly and any county-level characteristics that do not vary over time. The additional variables used by Cook and Ludwig (2006) in X_{it} are then meant to capture all other sources of variation that are correlated to both FSS and the log of of the homicide rate. Of course, one might worry that the set of controls included in X_{it} does not adequately capture remaining confounds after controlling for time and county effects.

We extend the analysis performed in Cook and Ludwig (2006) by allowing for a much larger set of potential control variables which may strengthen the plausibility of the claim that all sources of confounding variation have been captured. Specifically, we consider an essentially identical model

$$\log Y_{it} = \beta_0 + \beta_1 \log \text{FSS}_{it-1} + W_{it}'\beta_W + \alpha_i + \delta_t + \epsilon_{it}$$

which differs from (6.21) by our consideration of a large set of variables in W_{it}. We form W_{it} by taking variables compiled by the US Census Bureau. Basic variables include county-level measures of demographics, the age distribution, the income distribution, crime rates, federal spending, home ownership rates, house prices, educational attainment, voting paterns, employment statistics, and migration rates.[10] We note that W_{it} includes measures meant to capture all the variables controlled for in Cook and Ludwig (2006) in their X_{it}, though with our data and construction we do not reproduce their results exactly. However, we show below that we obtain similar results with both sets of variables. A key concern with the fixed effects model is that there is some feature of the counties that is correlated not just to the level of crime rates and gun ownership but also to the evolution of these variables. To flexibly allow for this possibility, we also include interactions of the initial (1980) values of all control variables with a linear, quadratic, and cubic term in time. With the main effects and interactions of initial conditions with a cubic trend, we end up with 978 total control variables.

While controlling for a large set of variables may make the assumption that all relevant confounds have been included in the model more plausible, including too many covariates may lower estimation precision and also complicates estimation of the variance of estimators as illustrated in Section 5.2. Using variable selection as outlined in this paper offers one potential resolution to this tension by allowing consideration of a large set of controls while maintaining parsimony and producing valid inferential statements under the assumption that the set of confounds that needs to be included after accounting for the full set of fixed effects is small relative to the sample size.

[10] The exact identities of the variables are available upon request. The entire dataset is taken from the U.S. Census Bureau USA Counties Database and can be downloaded at http://www.census.gov/support/USACdataDownloads.html.

We present estimation results in Table 3 with results for each dependent variable presented across the columns and each row corresponding to a different specification. As a baseline, we report numbers taken directly from the first row of Table 3 in Cook and Ludwig (2006) in the first row of Table 1 ("Cook and Ludwig (2006) Baseline"). Cook and Ludwig (2006) obtained these results by regressing log homicide rates on lagged log FSS, county and time fixed effects, and the baseline set of controls mentioned above and use these numbers as their baseline results.

We report results obtained from our data in Rows 2-4 (labeled "FSS + Census Baseline", " Full Set of Controls", and "Cluster Post-Double Selection").[11] In Row 2 of Table 3 ("FSS + Census Baseline"), we attempt to replicate the result from Row 1 using control variables gathered from the census that correspond to the variables indicated as being used in Cook and Ludwig (2006) Table 3, Row 1. Despite using slightly different data, we produce results that are fairly similar to those reported in Cook and Ludwig (2006). Specifically, Cook and Ludwig (2006) give point estimates (standard errors) of the coefficient on lagged log FSS of .086 (.038) for overall homicide rates and .173 (.049) for gun homicide rates; and we obtain estimated effects (standard errors) of .070 (.035) for overall homicide rates and of .178 (.046) for gun homicide rates. The discrepancy between the results is somewhat larger for non-gun homicide rates, though the results are still broadly consistent with each other. Cook and Ludwig (2006) report an estimated effect (standard error) of -.033 (.040) while we estimate the effect to be -.071 with a standard error of .038.

We provide the results based on the large set of controls in Rows 3 and 4 of Table 3. In Row 3, we present the results based on using all 978 potential controls in addition to the full set of county and time effects. Using all of the controls, the estimated effect of lagged suicide rates is small for each dependent variable. The estimated coefficients (standard errors) are only -.010 (.033) for overall homicide rates, .00004 (.044) for gun homicide rates, and -.033 (.042) for non-gun homicide rates. These results are relatively imprecise, and one could not rule out moderate sized positive or negative effects for any of the dependent variables. In addition, the simulation results illustrate that the estimated standard errors with a large number of controls may be inaccurate, suggesting that one should be hesitant in trusting these results as accurate standard errors may be even larger. Of course, it is not obvious that one would believe that all 978 controls are necessary though one may not be sure of the exact identities of the variables that should be included.

In Row 4, we present estimates of the effect of gun prevalence on homicide rates based on the post-double-selection method using Cluster-Lasso to select controls after partialing out the

[11] All results are based on weighted regression where we weight by the within-county average population over 1980-1999.

fixed effects. We also provide the identities of the selected controls in Table 4. For both overall homicide rates and gun homicide rates, the estimates based on Cluster-Lasso selected controls are very similar to those obtained with the baseline set of controls in our data though standard errors are slightly larger. For overall homicide results, the Cluster-Lasso estimate (standard error) is .079 (.043) compared to .070 (.035) with the baseline controls; and the Cluster-Lasso estimate (standard error) is .171 (.047) compared to .178 (.046) with the baseline controls when gun homicide is the dependent variable. This similarity is interesting given that the set of variables selected by Lasso differs substantively from the set of baseline controls. For the overall homicide rate, we would fail to reject the null hypothesis that gun prevalence as measured by suicide rates is not associated to homicide rates after controlling for a broad set of variables at the 5% level in the Cluster-Lasso results, though we would reject the hypothesis of no effect of gun prevalence on overall homicide rates at the 10% level. The result is stronger when the gun homicide rate is the dependent variable. In this case, one would draw the conclusion that more guns, as measured by the firearm suicide rate, is strongly positively associated with more homicides committed with firearms. Under the assumption that the set of controls considered is sufficient to account for relevant confounds, one could also take these estimated effects as causal. This assumption seems more plausible in the Cluster-Lasso results which allow for consideration of a richer set of controls than the baseline results.

Finally, we turn to the results with non-gun homicide as the dependent variable. In this case, there is a larger discrepancy between the baseline results and the results using controls selected by Cluster-Lasso, though one would draw the same qualitative conclusion in either case. With the baseline intuitively selected set of controls, the estimated effect of gun prevalence is -.071 with an estimated standard error of .038; and the estimated effect is smaller in magntitude, at -.019, with an estimated standard error of .040 using the Cluster-Lasso selected controls. In both cases, we would fail to reject the null hypothesis that gun prevalence as measured by suicide rates is not associated to non-gun homicide rates after controlling for a broad set of variables at conventional levels, and one could not rule out moderate positive or negative effects of gun prevalence on non-gun homicide at conventional levels using the Cluster-Lasso based results.

Overall, our Cluster-Lasso based results are broadly consistent with the claims of Cook and Ludwig (2006). We find a strong positive effect of gun prevalence on the firearm homicide rate after allowing for a large set of confounds and including a full set of county and time effects. We also find some evidence of a positive effect of gun prevalence on overall homicide rates, but produce an imprecise estimate of the effect on non-gun homicides which could be consistent with moderate positive or negative effects. The similarity to the Cook and Ludwig (2006) results adds further credibility to their claims as we allow for a richer set of confounding variables. We also note that the similarity between results following selection and results based on an intuitively selected initital set of controls is not mechanical as evidenced, for example, in the empirical example in Belloni, Chernozhukov, and Hansen (2014).

Appendix A. Cluster-Lasso Penalty Loadings Implementation

We organize implementation details for Cluster-Lasso and establish the asymptotic validity of the proposed algorithm in this appendix. Feasible options for setting the penalty level and the loadings for $j = 1, \dots, p$ are

$$\begin{array}{ll} \text{Initial:} & \widehat{\phi}_j = \sqrt{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{x}_{itj}\ddot{x}_{it'j}\ddot{y}_{itj}\ddot{y}_{it'j}}, \\ & \lambda = 2c\sqrt{nT}\Phi^{-1}(1-\gamma/(2p)), \\ \\ \text{Refined:} & \widehat{\phi}_j = \sqrt{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{x}_{itj}\ddot{x}_{it'j}\widehat{\epsilon}_{it}\widehat{\epsilon}_{it'}}, \\ & \lambda = 2c\sqrt{nT}\Phi^{-1}(1-\gamma/(2p)), \end{array} \tag{A.22}$$

where $c > 1$ is a constant, $\gamma \in (0, 1)$, and $\widehat{\epsilon}_{it}$ is an estimate of $\ddot{\epsilon}_{it}$. Let $K \geqslant 1$ denote a bounded number of iterations. We use $c = 1.1$, $\gamma = 0.1/\log(p \vee nT)$, and $K = 15$ in our empirical and simulation examples. In what follows, Lasso/Post-Lasso estimator indicates that the practitioner can apply either the Lasso or Post-Lasso estimator. Our preferred approach uses Post-Lasso at each step.

Algorithm of Cluster-Lasso penalty loadings

(1) Specify penalty loadings according to the initial option in (A.22). Use these penalty loadings in computing the Lasso/Post-Lasso estimator $\widehat{\beta}$ via equations (2.1) or (2.2). Then compute residuals $\widehat{\epsilon}_{it} = \ddot{y}_{it} - \ddot{x}'_{it}\widehat{\beta}$ for $i = 1, ..., n$ and $t = 1, ..., T$.

(2) If $K > 1$, update the penalty loadings according to the refined option in (A.22) and update the Lasso/Post-Lasso estimator $\widehat{\beta}$. Then compute a new set of residuals using the updated Lasso/Post-Lasso coefficients $\widehat{\epsilon}_{it} = \ddot{y}_{it} - \ddot{x}'_{it}\widehat{\beta}$ for $i = 1, ..., n$ and $t = 1, ..., T$.

(3) If $K > 2$, repeat step (2) $K - 2$ times. □

The algorithm above yields asymptotically valid penalty loadings in the sense that $\ell\phi_j \leqslant \widehat{\phi}_j \leqslant u\phi_j$ for every j with probability $1 - o(1)$, $\ell \xrightarrow{\mathrm{P}} 1$, and $u \leqslant C < \infty$. This fact is summarized in the following proposition. The proposition proceeds under an extended regularity condition:

Condition R′. (Extended Regularity)

(i) $\max_{1 \leqslant j \leqslant p} \left| \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\left(\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{y}_{it}\right)^2 - \mathrm{E}\left[\frac{1}{T}\left(\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{y}_{it}\right)^2\right]\right| / \mathrm{E}\phi_j^2 = o_{\mathrm{P}}(1)$.

(ii) $\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\left(\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{\epsilon}_{it}\right)^2\right|/\mathrm{E}\phi_j^2 = O_{\mathrm{P}}(1)$.

(iii) $\left(\max_{i,j,t}\ddot{x}_{itj}^2/\mathrm{E}\left[\phi_j^2\right]\right)\frac{s\log(p\vee nT)}{n\imath_T} = o_{\mathrm{P}}(1)$

Proposition 1 (Feasible Penalty Loadings)**.** *Under the conditions of Theorem 1 and Condition R', the penalty loadings $\widehat{\phi}_j$ constructed by the above algorithm are asymptotically valid. If $K \geqslant 2$, then $u \overset{\mathrm{P}}{\to} 1$.*

A.1. **Comments on Condition R'.** Condition R' imposes a set of high level conditions which could be verified under lower level primitive conditions. Condition R'(i) will generally be the most stringent as it may require convergence of a $\frac{1}{nT}$ normalized sum over nT^2 random elements which do not have mean zero. In the following two examples, we provide simple sample sets of sufficient primitive conditions under which Condition R' can be established. The first example covers a T fixed case as well as a case when $T \to \infty$ and data are strongly dependent in the sense that $\imath_T \propto 1$. The second case covers a scenario where $T \to \infty$ and $\imath_T \propto T$ which would be appropriate with weakly dependent data.

Example 1. *Suppose that T is fixed or that $\imath_T \propto 1$ and that $0 < m \leqslant \frac{\imath_T}{T}\mathrm{E}[\phi_j^2] \leqslant M < \infty$ for $1 \leqslant j \leqslant p$. Further, assume that the sequences of random variables $\{\ddot{y}_{it}, \ddot{x}_{it}\}_{t=1}^{T}$ are iid across i, that regressors are uniformly bounded with $\sup_{i,t,j}|\ddot{x}_{itj}| \leqslant B < \infty$, and that $\sup_t \mathrm{E}[\ddot{y}_{it}^4] \leqslant M < \infty$. Then Condition R'(i) is satisfied if $\frac{\log(p\vee n)^3}{n} \to 0$.*

The T fixed and $T \to \infty$ with $\imath_T \propto 1$ are similar in that essentially no information is accumulating in the time series dimension. In this case, rates of convergence are completely governed by the cross-sectional dimension and we see that Condition R'(i) may be satisfied when n grows quickly enough relative to $\log(p)$ under moment and boundedness conditions similar to those employed elsewhere in the literature, e.g. Example 3 in Belloni, Chernozhukov, and Hansen (2014).

Example 2. *Suppose that $T \to \infty$ with $\imath_T \propto T$ and that $0 < m \leqslant \frac{\imath_T}{T}\mathrm{E}[\phi_j^2] \leqslant M < \infty$ for $1 \leqslant j \leqslant p$. Suppose that $\{y_{it}, x_{it}\}$ is a strictly stationary strongly mixing (α-mixing) process with mixing coefficients satisfying $\theta(j) \leqslant \exp\{-2cj\}$.Further, suppose that the sequences of random variables $\{\{y_{it}, x_{it}\}_{t=1}^{T}, \alpha_i\}$ where α_i denotes unobserved individual specific heterogeneity are iid across i. Assume that observed random variables are uniformly bounded with $\sup_{i,t,j}|x_{itj}| \leqslant B$ and $\sup_{i,t}|y_{it}| \leqslant B$. Then Condition R'(i) is satisfied if $\frac{T\log(\max\{n,T,p\})^3}{n} \to 0$.*

Example 2 differs interestingly from Example 1 in requiring that $\frac{T}{n} \to 0$. The need for having n large relative to T in satisfying Condition R'(i) comes from the use of clustering even though the data is weakly dependent and the fact that $\mathrm{E}[x_{itj}y_{it}]$ will not generally be zero. The clustering estimator in the numerator then behaves like the variance of a strongly dependent process while the term in the denominator $\mathrm{E}[\phi_j^2]$ depends only on the $x_{itj}\epsilon_{it}$ process which is weakly dependent. Keeping the numerator from exploding relative to the denominator then requires a stronger condition on the rate of growth of T relative to n. Without this condition, one could not guarantee that ℓ and u (2.5) would remain bounded; specifically, one could produce $u \to \infty$ which could result in inflated initial penalty loadings and the failure to select any variables even when there are strong predictors among the set of variables considered. This feature suggests that there may be a price to pay in ability to select variables in this context in using the clustered variance estimator which is agnostic about dependence structures relative to a covariance estimator more tailored to a weakly dependent setting.

References

Altonji, J. G., and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053–1102.

Ando, T., and J. Bai (2015a): "Asset Pricing with a General Multifactor Structure," *Journal of Financial Econometrics*, 13, 556–604.

——— (2015b): "Panel Data Models with Grouped Factor Structure under Unknown Group Membership," *Journal of Applied Econometrics*, forthcoming.

Arellano, M. (1987): "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics*, 49(4), 431–434.

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, Arxiv, 2010.

Belloni, A., and V. Chernozhukov (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19(2), 521–547, ArXiv, 2009.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2014): "Program Evaluation with High-Dimensional Data," *arXiv:1311.2645*.

Belloni, A., V. Chernozhukov, and C. Hansen (2010): "LASSO Methods for Gaussian Instrumental Variables Models," 2010 arXiv:[math.ST], http://arxiv.org/abs/1012.1297.

——— (2013): "Inference for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III, 245–295.

——— (2014): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 119, 249–275.

Bester, C. A., and C. Hansen (2009): "Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model," *Journal of Business and Economic Statistics*, 27, 235–250.

——— (2014): "Grouped Effects Estimators in Fixed Effects Models," *Journal of Econometrics*, forthcoming.

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

Bonhomme, S., and E. Manresa (2013): "Grouped Patterns of Heterogeneity in Panel Data," working paper.

Candès, E., and T. Tao (2007): "The Dantzig selector: statistical estimation when p is much larger than n," *Ann. Statist.*, 35(6), 2313–2351.

Cattaneo, M., M. Jansson, and W. Newey (2010): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Working Paper, http://econ-www.mit.edu/files/6204.*

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," *Econometric Theory*, 28(1), 42–86.

Cook, P. J., and J. Ludwig (2006): "The social costs of gun ownership," *Journal of Public Economics*, 90, 379–391.

Duggan, M. (2001): "More Guns, More Crime," *Journal of Public Economics*, 109(5), 1086–1114.

Farrell, M. (2013): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," .

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

Hansen, C., and D. Kozbur (2014): "Instrumental variables estimation with many weak instruments using regularized JIVE," *Journal of Econometrics*, 182, 290–308.

Hansen, C. B. (2007): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics*, 141, 597–620.

Huang, J., J. L. Horowitz, and F. Wei (2010): "Variable selection in nonparametric additive models," *Ann. Statist.*, 38(4), 2282–2313.

Javanmard, A., and A. Montanari (2014): "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *arXiv:1306.3171v2.*

Jing, B.-Y., Q.-M. Shao, and Q. Wang (2003): "Self-normalized Cramr-type large deviations for independent random variables," *Ann. Probab.*, 31(4), 2167–2215.

Kock, A. B. (2014): "Oracle Inequalities for High-Dimensional Panel Data Models," Working Paper.

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

Lott, J. R. (2000): *More Guns, Less Crime, 2nd ed.* The University of Chicago Press, Chicago.

Meinshausen, N., and B. Yu (2009): "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, 37(1), 2246–2270.

Pötscher, B. M. (2009): "Confidence sets based on sparse estimators are necessarily large," *Sankhyā*, 71(1, Ser. A), 1–18.

Rudelson, M., and R. Vershynin (2008): "On sparse reconstruction from Fourier and Gaussian measurements," *Communications on Pure and Applied Mathematics*, 61, 10251045.

Rudelson, M., and S. Zhou (2011): "Reconstruction from anisotropic random measurements," *ArXiv:1106.1151.*

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

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

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014): "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *Annals of Statistics*, 42, 1166–1202.

Zhang, C.-H., and J. Huang (2008): "The sparsity and bias of the lasso selection in high-dimensional linear regression," *Ann. Statist.*, 36(4), 1567–1594.

Zhang, C.-H., and S. S. Zhang (2014): "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linaer Models," *Journal of the Royal Statistical Society: Series B*, 76, 217–242.

Table 1. Panel IV Simulations

	$p = n \times (T - 2)$				$p = n \times (T + 2)$			
	$n = 50$	$n = 100$	$n = 150$	$n = 200$	$n = 50$	$n = 100$	$n = 150$	$n = 200$
	Replications with No Instruments Selected							
Heteroscedastic Loadings	0	0	0	0	0	0	0	0
Clustered Loading	0	0	0	0	0	1	0	0
	A. Bias							
Oracle	-0.001	-0.007	0.000	-0.005	0.000	-0.008	0.000	-0.003
FE Oracle	-0.001	-0.010	-0.004	-0.004	0.000	-0.009	-0.004	-0.003
All	0.382	0.517	0.508	0.522				
Heteroscedastic Loadings	0.087	0.121	0.099	0.093	0.114	0.164	0.129	0.123
Clustered Loading	0.004	0.000	-0.001	-0.001	0.005	0.002	0.000	0.001
	B. RMSE							
Oracle	0.043	0.075	0.065	0.057	0.043	0.075	0.064	0.055
FE Oracle	0.077	0.078	0.060	0.053	0.076	0.076	0.060	0.053
All	0.384	0.518	0.508	0.522				
Heteroscedastic Loadings	0.120	0.151	0.120	0.111	0.143	0.186	0.146	0.140
Clustered Loading	0.081	0.078	0.062	0.053	0.081	0.075	0.061	0.054
	C. Size (Cluster s.e.)							
Oracle	0.067	0.048	0.057	0.052	0.064	0.048	0.053	0.048
FE Oracle	0.062	0.065	0.053	0.056	0.060	0.059	0.051	0.062
All	1.000	1.000	1.000	1.000				
Heteroscedastic Loadings	0.328	0.526	0.504	0.519	0.473	0.706	0.662	0.690
Clustered Loading	0.079	0.065	0.059	0.057	0.079	0.067	0.056	0.060
	D. Size (Heteroscedastic s.e.)							
Oracle	0.421	0.289	0.329	0.320	0.414	0.292	0.324	0.307
FE Oracle	0.249	0.240	0.234	0.214	0.246	0.222	0.238	0.214
All	1.000	1.000	1.000	1.000				
Heteroscedastic Loadings	0.586	0.710	0.709	0.715	0.680	0.848	0.824	0.836
Clustered Loading	0.275	0.247	0.236	0.221	0.274	0.227	0.234	0.221

This table presents simulation results for the high dimensional instrumental variables model with fixed effects. Estimators include our proposed Cluster-Lasso estimator (Clustered Loadings) and alternative estimators: heteroscedastic-Lasso (Heteroscedastic Loadings), 2SLS with all instruments (All), an oracle estimator that knows the values of the first-stage coefficients (Oracle), and an oracle estimator that knows the values of the first-stage coefficients and the fixed effects (FE Oracle). Bias, RMSE, and statistical size for 5% level tests using clustered standard errors and heteroscedastic standard errors are reported based on 1000 simulation replications.

TABLE 2. Panel Linear Model Simulations

	$p = n \times (T - 2)$				$p = n \times (T + 2)$			
	$n = 50$	$n = 100$	$n = 150$	$n = 200$	$n = 50$	$n = 100$	$n = 150$	$n = 200$
	A. Bias							
Oracle	0.003	0.002	0.001	0.000	-0.002	-0.002	0.000	0.001
FE Oracle	0.004	-0.002	0.001	0.001	0.000	-0.001	0.003	0.000
All	0.005	-0.001	-0.005	0.000				
Select over FE	0.070	0.014	-0.021	-0.020	0.075	0.011	-0.020	-0.019
Heteroscedastic Loadings	0.007	-0.023	-0.015	-0.012	-0.006	-0.034	-0.022	-0.020
Clustered Loading	0.040	0.006	0.010	0.009	0.035	0.007	0.011	0.008
	B. RMSE							
Oracle	0.089	0.058	0.051	0.042	0.084	0.060	0.050	0.042
FE Oracle	0.074	0.051	0.042	0.037	0.074	0.053	0.042	0.037
All	0.150	0.099	0.085	0.073				
Select over FE	0.108	0.087	0.052	0.046	0.109	0.089	0.054	0.045
Heteroscedastic Loadings	0.074	0.057	0.045	0.038	0.074	0.061	0.047	0.042
Clustered Loading	0.084	0.051	0.043	0.038	0.081	0.053	0.043	0.038
	C. Size (Cluster s.e.)							
Oracle	0.075	0.056	0.067	0.050	0.061	0.060	0.071	0.049
FE Oracle	0.060	0.052	0.047	0.057	0.060	0.062	0.046	0.058
All	0.514	0.467	0.494	0.458				
Select over FE	0.194	0.174	0.085	0.092	0.210	0.180	0.096	0.091
Heteroscedastic Loadings	0.085	0.101	0.076	0.081	0.088	0.151	0.119	0.127
Clustered Loading	0.093	0.062	0.059	0.057	0.093	0.071	0.066	0.062
	D. Size (Heteroscedastic s.e.)							
Oracle	0.330	0.289	0.329	0.297	0.311	0.316	0.322	0.301
FE Oracle	0.236	0.213	0.230	0.216	0.218	0.223	0.212	0.215
All	0.562	0.532	0.554	0.519				
Select over FE	0.414	0.328	0.310	0.313	0.448	0.355	0.303	0.290
Heteroscedastic Loadings	0.227	0.262	0.259	0.235	0.216	0.301	0.257	0.281
Clustered Loading	0.294	0.212	0.226	0.236	0.277	0.228	0.230	0.229

This table presents simulation results from a linear fixed effects model. Estimators include our proposed Cluster-Lasso estimator (Clustered Loadings), heteroscedastic-Lasso (Heteroscedastic Loadings), a double-selection estimator that includes the fixed effects in the set of variables to be selected over (Select over FE), fixed effects using all controls (All), an oracle estimator that knows the values of the coefficients on the control variables (Oracle), and an oracle estimator that knows the values of the coefficients on the controls variables and the fixed effects (FE Oracle). Bias, RMSE, and statistical size for 5% level tests using clustered standard errors and heteroscedastic standard errors are reported based on 1000 simulation replications.

TABLE 3. Estimates of the Effect of Gun Prevalence on Homicide Rates

	Overall	Gun	non-Gun
Cook and Ludwig (2006) Baseline	0.086 (0.038)	0.173 (0.049)	-0.033 (0.040)
FSS + Census Baseline	0.070 (0.035)	0.178 (0.046)	-0.071 (0.038)
Full Set of Controls	-0.010 (0.033)	0.000 (0.044)	-0.033 (0.042)
Cluster Post-Double Selection	0.079 (0.043)	0.171 (0.047)	-0.019 (0.040)

This table presents estimates of the effect of gun ownership on homicide rates for a panel of 195 US Counties over the years 1980-1999. The columns "Overall", "Gun", and "non-Gun" respectively report the estimated effect of gun prevalence on the log of the overall homicide rate, the log of the gun homicide rate, and the log of the non-gun homicide rate. Each row corresponds to a different specficiation as described in the text. In each specification, the outcome corresponding to the column label is regressed on lagged log(FSS) (a proxy for gun ownership) and additional covariates as described in the text. Each specification includes a full set of year and county fixed effects. Standard errors clustered by county are provided in parentheses.

TABLE 4. Variables Selected

A. log(FSS)
Owner occupied housing units
Renter occupied housing units
Males 15 yrs widowed
Institutionalized population
$t\times$ (Total bank deposits)$_0$
$t\times$ (% Change in households)$_0$
B. Overall homicide
Persons 5 yrs and over by residence - Same house for last 5 yrs
Vote cast for president, third party candidate
$t^3\times$ (Valuation of new housing by building permits)$_0$
C. Gun homicide
Resident population age 50 - 54 years
Vote cast for president, third party candidate
Owner occupied housing units
Families with income 15,000 - 19,999
D. non-Gun homicide
Resident population median age
Persons per household
$t\times$ (Hispanic persons 25 years and over)$_0$

The table presents selected variables by Cluster-Lasso in the gun example using our extended list of controls variables. Variables selected with lagged $\log(FSS)$, the log of the overall homicide rate, the log of the gun homicide rate, and the log of the non-gun homicide rate are given in Panels A, B, C, and D respectively.