The role of sparsity in panel data models*

Andrew Adrian Yu Pua†

October 26, 2015

Abstract

Recent research in panel data has focused on making sense of what we mean by unobserved heterogeneity. In the context of individual-specific fixed effects, sparsity can be thought of as filling out a gray area between full heterogeneity and full homogeneity in regression coefficients. In this paper, sparsity of the fixed effects means that some fixed effects have absolute value of zero while others are bounded and other are large. The proposed estimator attempts to detect the large values of the fixed effects so that they can be removed in the second step. In particular, I construct a panel lasso estimator derived from a least squares criterion with an L_1-penalty on the fixed effects only. I tune the regularization parameter to encourage sparsity and allow for contemporaneous exogenous regressors. As a second step, I remove the large non-zero fixed effects so that pooled OLS may be applied. I justify the use of these estimators using fixed-T asymptotics. Monte Carlo results indicate good performance even when both n and T are small as long as the sparsity assumption is true. As an empirical application, I revisit the state-level data constructed by van der Weide and Milanovic (2014) to demonstrate the negative impact of inequality on income growth in the United States. I show how this effect has been changing for the better over time by applying the panel lasso to a dynamic model with two time periods.

Keywords: Lasso, Incidental parameters, Pooled OLS, Dynamic panel data
JEL Classification: C23, C55

*I thank Maurice Bun and Sébastien van Bellegem for their extensive comments on an earlier draft. I thank Roy van der Weide for allowing me access to the data that I analyze in this paper. Comments from Pavel Čížek, Jianqing Fan, Artūras Juodis, Frank Kleibergen, Renata Rabovič are greatly appreciated. I also thank participants of the 4th Amsterdam-Bonn Econometrics Workshop, the CORE-ILSM Lectures on High Dimensional Econometrics, and the 2015 Netherlands Econometric Study Group. Finally, parts of the paper were written while I was being financially supported by the European Commission in the framework of the European Doctorate in Economics – Erasmus Mundus (EDEEM). All errors remain mine. Updated versions, if available, can be found in http://andrew-pua.ghost.io.

†Faculty of Economics and Business, University of Amsterdam, andrewypua@gmail.com

1 Introduction

We have seen increased collection of longitudinal or panel data through active or passive means in recent years. We can study these repeated measurements in three ways – (a) analyze the repeated measurements for each cross-sectional unit separately, (b) analyze the cross-sectional information, and (c) pool the both cross-sectional and time-series information together.

Methods in time series analysis can be used in situation (a) but will only be feasible when the number of repeated measurements is sufficiently large. The latter case precludes studying panels with a short time series dimension, typically collected for purposes of crafting policy. Methods in cross-sectional analysis can be used in situation (b) but precludes the study of the dynamics of change unless the time series dimension is also large. A compromise would then be to use methods that accomplish (c).

Unfortunately, there is much leeway as to how we should pool information available in panel data. Traditionally, econometricians have introduced cross-sectional heterogeneity in the parameters of a panel data model. Research during the 1960s up to the 1980s, cross-sectional heterogeneity is usually accomplished via the variance components model and the random coefficients model. These models typically impose parametric assumptions on the distribution of heterogeneity so that the dimension of the parameter space can be reduced substantially. Recent research has been aimed at completely removing these parametric assumptions. Success in this area has been mixed but a lot of progress has been made.

In particular, recent results have been negative with respect to fixed-T identification and fixed-T consistent estimation (see the most recent survey by Arellano and Bonhomme (2011)). However, a major insight behind recent results is the need for reducing the support of the fixed effects relative to the support of the dependent variable. Bonhomme (2012) show how this reduction in the support aids in constructing moment conditions for the structural parameters. Despite these negative results, Browning and Carro (2010) argue that we have actually not allowed for full heterogeneity at all. In particular, they argue for a fully heterogeneous setup where slope coefficients are allowed to vary across observations but still be time-invariant. Another way to interpret heterogeneity is to allow for time-invariant heterogeneity in the inverse link functions (and not the coefficients of the linear predictor) for single-index panel data models as proposed by Chen, Gao, and Li (2013).

Notice that the previous descriptions of heterogeneity assume that cross-sectional units are totally different from one another. At the other extreme, all cross-sectional units are assumed to be the same (with respect to model parameters). There is a large middle ground that needs to be explored. Grouping and clustering methods come to mind because they allow the data to determine which units can be pooled and which cannot. Furthermore, partial pooling allows for a possibility to implement the reduction in the support of the fixed effects. Recent research on grouped heterogeneity by Bonhomme and Manresa (2015)

point toward this possibility. They even allow the grouping to vary over time. Yet another way to implement partial pooling is proposed by Sarafidis and Weber (2011) where they allow for an unknown number of clusters in the data and full homogeneity is assumed within each cluster.

In this chapter, I argue that sparsity may be a useful device to accomplish a reduction in the support of the fixed effects and to allow the data to determine the groups that may be present in the data. In particular, there are economic and empirical situations for which some cross-sectional units can have the same value for the individual-specific effect. For instance, an econometric method should be able to accommodate the situation where only a subset of units obey conditional moment restrictions implied by an economic model. This is where we must account for partial pooling and where a sparsity assumption on the individual-specific effects can be a useful technical device. Furthermore, it is of interest to try to identify these deviations in the same manner in which we want to be able to detect outliers to obtain some form of robustness.

Recent work by Fan, Tang, and Shi (2012) indicate that it is possible to estimate the structural parameter of a linear model with exogenous covariates with just $T = 1$ despite allowing for the intercept to vary across observations. Their idea was to divide the incidental parameters into three types – those that are very large that they can be treated as outliers, those that are zero, and those that are non-zero but small enough that they can be treated as zero asymptotically. I show how to extend their arguments to the linear panel data case but allowing for contemporaneously exogenous variables. I also modify their procedure for selecting the data-driven regularization parameter. Unfortunately, not all the results in Fan, Tang, and Shi (2012) survive the extension as we will see in the next section.

Although sparsity has been used in machine learning and big data situations, the focus has always been settings where the number of covariates is extremely large relative to the sample size. I restrict myself to the setting where the regressor vector is still finite-dimensional. In contrast, Kock (2013) differences out the incidental parameters first before proposing a penalty method for the differenced model and allows the regressor vector to be high-dimensional. Kock (2014) extends the previous paper to allow the possibility that the incidental parameters are weakly sparse. In his context, weak sparsity means that the L_1 norm of all the incidental parameters is small. As a result, the values of the incidental parameters need not be zero at all. In contrast, I explicitly have zero-valued incidental parameters but allow for some of these parameters to be small enough that they can be taken as zero asymptotically. Furthermore, these two papers by Kock are confined to regressors that are strictly exogenous. Kock and Tang (2014) extends these papers further to dynamic panels and allow for predetermined regressors. All these developments are under a framework where n and T are allowed to vary. Furthermore, their results are in the form of oracle inequalities. These inequalities provide upper bounds for the estimation error (in some suitable norm) as a function of the design matrix and the dimensions of the problem.

In contrast, my modifications to Fan, Tang, and Shi (2012) for the panel data case allow me to consider contemporaneously exogenous regressors and a fixed number of time periods T. I introduce these modifications and the resulting consequences in Section 2. I use Monte Carlo simulations to study the finite sample performance of the two-step panel lasso estimator in Section 3. I revisit the relationship between inequality and growth using the microdata collected by van der Weide and Milanovic (2014). I end with some concluding remarks, suggestions for future research, and a technical appendix containing some proofs of the main results.

2 Panel lasso for the linear model

2.1 Setup and notation

Consider the data generating process where

$$y_{it} = \alpha_{i0} + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}_0 + \epsilon_{it}, \;\; i = 1, \ldots, n; \; t = 1, \ldots, T \tag{1}$$

where $\left(\alpha_{10}, \alpha_{20}, \ldots, \alpha_{N0}, \boldsymbol{\beta}_0\right)$ are the true values of the parameters and $\boldsymbol{\beta}_0 \in \mathbb{R}^d$. In contrast to the machine learning literature, I assume that d is fixed and does not grow with sample size. Define the averages $\bar{\mathbf{x}}_i = T^{-1} \sum_{t=1}^{T} \mathbf{x}_{it}$ and $\bar{\epsilon}_i = T^{-1} \sum_{t=1}^{T} \epsilon_{it}$. Let $a_+ = \max\{a, 0\}$ be the positive part of a, $\operatorname{sgn}(a)$ be the sign function, and $\|\cdot\|_2$ be the L_2-norm. Let $B_C\left(\boldsymbol{\beta}_0\right) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \left|\beta_j - \beta_{j0}\right| \leq C, \; 1 \leq j \leq d\}$ for some constant $C > 0$.

I impose the following assumptions:

A1 (Independence) The errors ϵ_{it} are independent across i.

A2-1 (Predeterminedness) The errors ϵ_{it} and the covariates $\mathbf{x}_i^t = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it})$ satisfy $\mathbb{E}(\epsilon_{it} \mid \mathbf{x}_i^t) = 0$ for all i and t.

A2-2 (Contemporaneous exogeneity) The errors ϵ_{it} and the covariates $\mathbf{x}_{it}$ satisfy $\mathbb{E}(\epsilon_{it} \mid \mathbf{x}_{it}) = 0$ for all i and t.

A3 (Behavior of averages) Assume that $\mathbb{E}\left(\|\bar{\mathbf{x}}_i\|_2\right) < \infty$ and $\mathbb{E}(\bar{\epsilon}_i) < \infty$. There exists $\kappa_n, \gamma_n = o\left(\sqrt{n}\right)$ such that, as $n \to \infty$, we have

$$\Pr\left(\max_{1 \leq i \leq n} \|\bar{\mathbf{x}}_i\|_2 > \kappa_n\right) \;\; \to \;\; 0, \tag{2}$$

$$\Pr\left(\max_{1 \leq i \leq n} |\bar{\epsilon}_i| > \gamma_n\right) \;\; \to \;\; 0. \tag{3}$$

A4 (Sparsity) Each i belongs to one and only one of the three possible index sets $\{1, \ldots, s_1\}$, $\{s_1 + 1, \ldots, s\}$, and $\{s + 1, \ldots, n\}$. If $i \in \{1, \ldots, s_1\}$, then $\kappa_n = o\left(|\alpha_{i0}|\right), \gamma_n = o\left(|\alpha_{i0}|\right)$. If $i \in \{s_1 + 1, \ldots, s\}$, then $|\alpha_{i0}| < \gamma_n$. If $i \in \{s + 1, \ldots, n\}$, then $\alpha_{i0} = 0$.

Assumption A1 is a standard assumption imposed in panel data models without cross-sectional dependence. Assumptions A2-1 or A2-2 allow us to consider dynamics or feedback effects. Implementations of GMM estimators for panel data models usually maintain Assumption A2-1 (see Bun and Sarafidis (2015) for a survey). Assumption A2-2 is usually imposed in pooled OLS (see Wooldridge (2010)). Fan, Tang, and Shi (2012) impose assumptions on the behavior of the covariates and the errors similar to A3. The difference is that we impose tail behavior assumptions on the time series averages for every i rather than on the individual values. The existence of $\kappa_n, \gamma_n = o\left(\sqrt{n}\right)$ is guaranteed by A1 and Markov's inequality. Finally, there are three types of incidental parameters by A4 – s_1 of them are "large" incidental parameters, $s - s_1$ of them are bounded, and $n - s$ of them are zero. Note that A4 imposes an assumption on the number and the size of the incidental parameters. If we can detect which of the cross-sectional units are zero, then these units can now be pooled together to recover a consistent estimator for $\boldsymbol{\beta}_0$. Note that under large-n asymptotics, the number of each type of incidental parameter may grow with n. We will see later how the growth in the number of each type of incidental parameter has to be restricted so that consistency and asymptotic normality would be obtained.

Under what circumstances would it be plausible for assumption A4 to hold? Consider the following linear model where

$$y_{it} = \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0 + \omega_{it}, \quad i = 1, \ldots, n; \; t = 1, \ldots, T. \tag{4}$$

Decompose ω_{it} into $\mathbb{E}(\omega_{it}|\mathbf{x}_{it})$ and its residual $\omega_{it} - \mathbb{E}(\omega_{it}|\mathbf{x}_{it})$. Let $\alpha_{i0} = \mathbb{E}(\omega_{it}|\mathbf{x}_{it})$ be the portion of the error ω_{it} representing some model deficiency specific to the ith unit that is correlated with the included regressors. Let the residual $\omega_{it} - \mathbb{E}(\omega_{it}|\mathbf{x}_{it})$ be equal to ϵ_{it}. We have now produced (1) that can potentially satisfy the assumptions laid out above from (4). Therefore, the units for which $\alpha_{i0} = 0$ can represent the units for which the conditional moment restriction $\mathbb{E}(y_{it}|\mathbf{x}_{it}) = \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0$ is appropriate. The units for which α_{i0} are close enough to zero may be treated as zero asymptotically using the proposed panel lasso estimator. It then becomes important to detect the units for which there is some serious model deficiency. The described setting may also apply to situations where we have endogenous regressors but are unable to find valid instruments. Think of α_{i0} as the unit-specific correlation between the error ω_{it} and $\mathbf{x}_{it}$. It is possible that only a subset of the units have a regressor vector that is endogenous. It is therefore of interest to detect these units so that we are still able to consistently estimate $\boldsymbol{\beta}_0$ after removing these units from the sample.

2.2 Estimation and inference

To develop an estimator for $\boldsymbol{\beta}_0$, consider minimizing the least squares objective function subject to an L_1-penalty[1] on the incidental parameters, i.e.

$$\min_{(\alpha_1,\alpha_2,\ldots,\alpha_n,\boldsymbol{\beta})} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}-\alpha_i-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)^2+\sum_{i=1}^{n}2\lambda|\alpha_i|, \tag{5}$$

where $\lambda \geq 0$ is some user-specified regularization parameter. This parameter takes on nonnegative values and governs the rate at which shrinkage toward zero is being applied to each of the α_i. Large values of λ will tend to shrink the α_i's toward zero. Therefore, the minimizer of (5) is the pooled OLS estimator when $\lambda \to \infty$. On the other hand, we obtain the within estimator when $\lambda \to 0$.

A minimizer $(\alpha_1, \alpha_2, \ldots, \alpha_n, \boldsymbol{\beta})$ of (5) satisfies the following first-order conditions:[2]

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}-\alpha_i-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\mathbf{x}_{it} = 0, \tag{6}$$

$$\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\alpha_i-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)-\lambda\frac{\alpha_i}{|\alpha_i|} = 0. \tag{7}$$

For an arbitrary $\boldsymbol{\beta}$, we can solve for α_i from (7) as

$$\begin{cases} \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\left(y_{it}-\widehat{\alpha_i}(\boldsymbol{\beta})-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)=\lambda\operatorname{sgn}\left(\widehat{\alpha_i}(\boldsymbol{\beta})\right) & \text{if } \widehat{\alpha_i}(\boldsymbol{\beta})\neq 0, \\ \left|\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right|-\lambda\leq 0 & \text{if } \widehat{\alpha_i}(\boldsymbol{\beta})=0. \end{cases} \tag{8}$$

(8) can be rewritten as a soft-threshold estimator, i.e.

$$\widehat{\alpha_i}(\boldsymbol{\beta})=\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right|-\lambda\right)_{+}\operatorname{sgn}\left(\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right). \tag{9}$$

Substituting this into (6) gives a profiled estimating function for $\boldsymbol{\beta}$:

$$g(\boldsymbol{\beta})=\left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}^{\mathrm{T}}\right)\boldsymbol{\beta}-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\left(y_{it}-\widehat{\alpha_i}(\boldsymbol{\beta})\right) \tag{10}$$

The panel lasso estimator for $\boldsymbol{\beta}_0$, denoted by $\widehat{\boldsymbol{\beta}}$, solves $g\left(\widehat{\boldsymbol{\beta}}\right)=0$. Let $\widehat{\alpha_i}=\widehat{\alpha_i}\left(\widehat{\boldsymbol{\beta}}\right)$.

Since the objective is to derive the asymptotic properties of $\widehat{\boldsymbol{\beta}}$, we need to determine how

[1] Imposing an L_2-penalty leads to ridge regression. I do not use this penalty because I am working with Assumption A4. The L_2-penalty only shrinks estimators toward zero.

[2] To get the derivative of the absolute value function $|\alpha|$, note that $|\alpha|=\sqrt{\alpha^2}$. So, $\partial_\alpha|\alpha|=\partial_\alpha\sqrt{\alpha^2}=\left(\alpha^2\right)^{-1/2}\times 2\alpha/2=\alpha/|\alpha|$ provided that $\alpha\neq 0$.

well (9) classifies the ith unit into one of the sets $\{1,\ldots,s_1\}$, $\{s_1+1,\ldots,s\}$, and $\{s+1,\ldots,N\}$. Note that (9) depends on the signs of $\left|\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right|-\lambda$ and $\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)$. Substitute the model into the preceding expressions and define the following index sets:

$$\begin{aligned}
S_{10} &= \left\{s+1\leq i\leq n:\left|\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\boldsymbol{\beta}\right)+\bar{\epsilon}_i\right|\leq\lambda\right\},\\
S_{11} &= \left\{1\leq i\leq s_1:\left|\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\boldsymbol{\beta}\right)+\bar{\epsilon}_i\right|\leq\lambda\right\},\\
S_{12} &= \left\{s_1+1\leq i\leq s:\left|\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\boldsymbol{\beta}\right)+\bar{\epsilon}_i\right|\leq\lambda\right\}.
\end{aligned}$$

Call S_{20}, S_{21}, and S_{22} the sets where we drop the absolute values and replace $\leq\lambda$ with $>\lambda$ in the definitions of S_{10}, S_{11}, and S_{12} respectively. Finally, call S_{30}, S_{31}, and S_{32} the sets where we drop the absolute values and replace $\leq\lambda$ with $<-\lambda$ in the definitions of S_{10}, S_{11}, and S_{12} respectively. By the definitions above, S_{10}, S_{20}, and S_{30} are mutually disjoint. The same applies to S_{11}, S_{21}, and S_{31} and S_{12}, S_{22}, and S_{32}. By assumption A4, $\alpha_{i0}=0$ for all $i\in S_{10},S_{20},S_{30}$. Note that these index sets will depend on $\boldsymbol{\beta}$. For an arbitrary index set S, I use $\widehat{S}$ to denote the result if we plug in the panel lasso estimator $\widehat{\boldsymbol{\beta}}$ into S.

To analyze whether the panel lasso estimator is consistent, I have to analyze the components of the estimating equation $g\left(\widehat{\boldsymbol{\beta}}\right)=0$ after substituting (1) into (10):

$$\mathbf{W}_n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)=\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\alpha_{i0}+\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\epsilon_{it}-\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\widehat{\alpha_i}, \tag{11}$$

where

$$\mathbf{W}_n=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}^{\mathrm{T}}.$$

The last term of (11) can be rewritten depending on which set i belongs, i.e.

$$\sum_{i\in S}\bar{\mathbf{x}}_i\widehat{\alpha_i} = \begin{cases}
\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i-\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{20}}\\
\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\alpha_{i0}+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i-\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{21}},\ \widehat{S_{22}}\\
\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i+\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{30}}\\
\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\alpha_{i0}+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i+\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{31}},\ \widehat{S_{32}}\\
\mathbf{0} & \text{if } S=\widehat{S_{10}},\ \widehat{S_{11}},\ \widehat{S_{12}}
\end{cases} \tag{12}$$

To further simplify (11), we need to say something about the contents of the index sets defined earlier as $n\to\infty$. But then, we would have to specify how to tune the regularization parameter, i.e. I assume that

$$\kappa_n=o\left(\lambda\right),\ \alpha\gamma_n\leq\lambda,\ \lambda=o\left(\min\left\{\alpha^*,\sqrt{n}\right\}\right), \tag{13}$$

where $\alpha > 2$ and $\alpha^* = \min_{1 \leq i \leq s_1} \left|\alpha_i^*\right|$. This means that λ should be large enough to overrule the tail behavior of the time series averages of the regressors and the idiosyncratic error but small enough that it does not overrule the smallest of the "large" incidental parameters. As a result, I am able to extend Lemma 3.1 of Fan, Tang, and Shi (2012) to the panel data case and I present the details of the proof in the appendix.

Lemma 1 (Contents of index sets)**.** *Assume that A1, A2-1 (or A2-2), A3, and A4 hold. Let $n \to \infty$. For every $C > 0$ and for every $\boldsymbol{\beta} \in B_C\left(\boldsymbol{\beta}_0\right)$, with probability going to 1,*

$$\begin{array}{lll} S_{10} = S_{10}^*, & S_{11} = \varnothing, & S_{12} = S_{12}^* \\ S_{20} = \varnothing, & S_{21} = S_{21}^*, & S_{22} = \varnothing \\ S_{30} = \varnothing, & S_{31} = S_{31}^*, & S_{32} = \varnothing \end{array}$$

where $S_{10}^* = \{s+1, s+2, \ldots, n\}$, $S_{12}^* = \{s_1+1, s_1+2, \ldots, s\}$, $S_{21}^* = \{1 \leq i \leq s_1 : \alpha_{i0} > 0\}$ and $S_{31}^* = \{1 \leq i \leq s_1 : \alpha_{i0} < 0\}$.

Notice that the preceding lemma enables us to allocate the indices $\{1, \ldots, n\}$ into four sets asymptotically – (a) S_{10}^* contain the indices for the units whose incidental parameter values are equal to zero, (b) S_{12}^* contain the indices for the units whose incidental parameter values are "bounded", (c) S_{21}^* contain the indices for the units whose incidental parameter values are "large" and positive, and (d) S_{31}^* contain the indices for the units whose incidental parameter values are "large" and negative. However, this result is not enough to guarantee consistency of the panel lasso estimator for $\boldsymbol{\beta}_0$.

Notice that the left hand side of (11) still contains terms that do not disappear in the limit unless we impose additional restrictions on the rate of growth of the number of the " bounded" and "large" incidental parameters. The proof of the next theorem can be found in the appendix.

Theorem 1 (Consistency of the panel lasso estimator)**.** *Assume that A1, A2-1 (or A2-2), A3, and A4 hold. Further assume that (i) $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$, where $\mathbf{W}$ is nonsingular, (ii) $s - s_1 = o\left(n/\left(\kappa_n \gamma_n\right)\right)$, (iii) λ obeys (13), and (iv) $s_1 = O(1)$. Then, for some $\overline{C} > 0$, wpg 1, there exists a unique estimator $\widehat{\boldsymbol{\beta}} \in B_{\overline{C}}\left(\boldsymbol{\beta}_0\right)$ such that $g\left(\widehat{\boldsymbol{\beta}}\right) = 0$ hold and $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.*

The theorem provides us with an alternative consistent estimator for the linear dynamic panel data model (possibly with feedback effects) but it would require a sparsity assumption. This theorem also differs from Theorem 3.2 of Fan, Tang, and Shi (2012) because (iv) is present. This condition bounds the number of "large" incidental parameters by a constant even as $n \to \infty$. However, it buys us the possibility to include predetermined and even contemporaneously exogenous variables and still be able to obtain a consistent estimator. Furthermore, Fan, Tang, and Shi (2012) assume a zero mean for the covariates. I do not impose that assumption at all. Had we imposed this assumption, we can use a less restrictive condition on the number of "large" incidental parameters, i.e., $s_1 = o\left(n/\left(\kappa_n \gamma_n\right)\right)$.

I now construct a two-step estimator. First, define the following events:

$$\begin{aligned}\mathscr{E}_1 &= \{\widehat{\alpha_i} \neq 0 \text{ for } i = 1, \ldots, s_1\}, \\ \mathscr{E}_2 &= \{\widehat{\alpha_i} = 0 \text{ for } i = s_1 + 1, \ldots, s, s+1, \ldots, n\}.\end{aligned}$$

The next lemma allows us to construct a two-step estimator by choosing the subset of the n units whose α_{i0} was estimated to be $\widehat{\alpha_i} = 0$. As long as we have a consistent estimator for $\boldsymbol{\beta}_0$, we would be able to detect the indices of the units who have "large" incidental parameters with high probability. Unfortunately, the lemma states that we are unable to estimate their values consistently. More importantly, we are able to detect the zero-valued incidental parameters correctly but we shrink all the bounded incidental parameters to zero. The proof is available in the appendix.

Lemma 2 (Partial consistency)**.** *Let* $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2$*. If* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$*, then* $\Pr(\mathscr{E}) \to 1$ *under A1 to A4.*

In principle, the lemma applies to any initial consistent estimator of $\boldsymbol{\beta}_0$, even those that do not explicity encourage sparsity. An example would be the usual GMM estimator. Unfortunately, the GMM estimator is not available under Assumption A2-2.

We now reestimate (1) using the data from the subset for which $\widehat{\alpha_i} = 0$ using this lemma. Define

$$\widehat{I_0} = \{1 \leq i \leq n : \widehat{\alpha_i} = 0\}$$

to be the subset under consideration. The two-step panel lasso estimator $\widetilde{\boldsymbol{\beta}}$ can now be defined as the minimizer of

$$\min_{\boldsymbol{\beta}} \frac{1}{nT} \sum_{i \in \widehat{I_0}} \sum_{t=1}^{T} \left(y_{it} - \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)^2, \tag{14}$$

Notice that (14) is exactly the least squares objective function restricted to observations belonging to $\widehat{I_0}$. Define the following matrices for $i \in \widehat{I_0}$:

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1}^{\mathrm{T}} \\ \mathbf{x}_{i2}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{iT}^{\mathrm{T}} \end{pmatrix}, \ \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix}, \ \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT} \end{pmatrix}$$

The solution to (14) is given by the usual pooled least squares estimator, i.e.,

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{i \in \widehat{I_0}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_i\right)^{-1} \left(\sum_{i \in \widehat{I_0}} \mathbf{X}_i^{\mathrm{T}} \mathbf{y}_i\right). \tag{15}$$

The next theorem shows that two-step panel lasso estimator $\widetilde{\boldsymbol{\beta}}$ is consistent as $n \to \infty$. The underlying idea is to use the previous lemma and apply the system OLS consistency theorem (Theorem 7.1) of Wooldridge (2010), along with an assumption on the rate of growth of the number of bounded incidental parameters.

To be specific, we can rewrite (15) as

$$
\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right) \\
&= \left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right) + \left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right)\left(\Pr\left\{\widehat{I_0} \neq I_0\right\} + \Pr\left\{\widehat{I_0} = I_0\right\}\right) \\
&\quad - \left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right)\left(\Pr\left\{\widehat{I_0} \neq I_0\right\} + \Pr\left\{\widehat{I_0} = I_0\right\}\right) \\
&= \left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right) + \underbrace{\left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right)\Pr\left\{\widehat{I_0} \neq I_0\right\}}_{R_1} \\
&\quad \underbrace{- \left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right)\Pr\left\{I_0 \neq \widehat{I_0}\right\}}_{R_2} \\
&\xrightarrow{p} \boldsymbol{\beta}_0 + \left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\iota_T\alpha_{i0} + \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i\right),
\end{aligned}
$$

where ι_T is a $T\times 1$ vector of ones. The terms R_1 and R_2 involve $\Pr\left\{\widehat{I_0} \neq I_0\right\}$, which converges to 0 as $n \to \infty$ by Lemma 2. The term involving the incidental parameters can be evaluated as follows:

$$
\left\|\frac{1}{n}\sum_{i\in I_0} \mathbf{X}_i^{\mathrm{T}}\iota_T\alpha_{i0}\right\|_2 = \left\|\frac{1}{n}\sum_{i\in I_0} T\bar{\mathbf{x}}_i\alpha_{i0}\right\|_2 \leq \frac{T}{n}\sum_{i=s_1+1}^{s} \|\bar{\mathbf{x}}_i\|_2 |\alpha_{i0}| \leq T\left(\frac{s-s_1}{n}\right)\kappa_n\gamma_n.
$$

Here we used assumption A4 to show that $\alpha_{i0} = 0$ for $i = s+1,\ldots,n$. As a result, the term becomes $o_p(1)$ when $s - s_1 = o(n/(\kappa_n\gamma_n))$. Take note that T is taken as fixed here. The term involving the errors has probability limit equal to zero because of A2-2. As a result, we have the following theorem:

Theorem 2 (Consistency of the two-step panel lasso estimator)**.** *Suppose that the conditions in Theorem 1 hold. Further assume that (i)* $\mathbf{A} = E\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{X}_i\right)$ *is nonsingular and (ii)* $s - s_1 = o(n/(\kappa_n\gamma_n))$*. Then* $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$*.*

I now show the asymptotic normality of the two-step panel lasso estimator. Note that $\Pr\left(\sqrt{n}R_1 = 0\right) \leq \Pr\left\{\widehat{I_0} \neq I_0\right\} = 0$. Analogously, we have $\Pr\left(\sqrt{n}R_2 = 0\right) = 0$. Consider

again the argument earlier for consistency. The term involving the incidental parameters will only be $o_p(1)$ when $s - s_1 = o\left(\sqrt{n}/(\kappa_n \gamma_n)\right)$ to asymptotically remove the influence of the bounded incidental parameters. A central limit theorem can then be applied to the term involving the errors. The result then follows from OLS asymptotic normality theorem for systems of equations (Theorem 7.2) in Wooldridge (2010).

Theorem 3 (Asymptotic normality of the two-step panel lasso estimator)**.** *Suppose that the conditions in Theorem 1 hold. Further assume that (i)* $\mathbf{A} = E\left(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{X}_i\right)$ *is nonsingular and (ii)* $s - s_1 = o\left(\sqrt{n}/(\kappa_n \gamma_n)\right)$*. Then* $\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N\left(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right)$*, where* $\mathbf{B} = Var\left(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\epsilon}_i\right)$*.*

Following standard arguments like those in Wooldridge (2010), a consistent estimator of the asymptotic variance of $\widetilde{\boldsymbol{\beta}}$ is given by

$$\widehat{\mathbf{V}} = \left(\sum_{i \in \widehat{I_0}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_i\right)^{-1} \left(\sum_{i \in \widehat{I_0}} \mathbf{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{\epsilon}}_i^{\mathrm{T}} \mathbf{X}_i\right) \left(\sum_{i \in \widehat{I_0}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_i\right)^{-1},$$

where $\widehat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \mathbf{X}_i \widetilde{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_i - \mathbf{X}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ is a consistent estimator of $\boldsymbol{\epsilon}_i$. Note that this estimator is a robust variance estimator which allows for arbitrary serial correlation and time-series heteroscedasticity.

2.3 Choice of regularization parameter

So far we only have a theoretical specification for the regularization parameter λ as seen in (13). In practice, the choice of λ would have to be data-driven. A feasible procedure for the proposed method is as follows:

1. Apply OLS to the model where $y_{i1} = \mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_{i1}$ for $i = 1, \ldots, n$. Obtain the residuals from the resulting regression, i.e. $\widehat{\epsilon_{i1}} = y_{i1} - \mathbf{x}_{i1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}^{OLS}}$ for every i.

2. Select n_{pure} observations that correspond to the smallest values of $\{|\widehat{\epsilon_{11}}|, \ldots, |\widehat{\epsilon_{n1}}|\}$.

3. Apply OLS once again to the model where $y_{i1} = \mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_{i1}$ but only for the selected n_{pure} observations. Obtain the new set of residuals $\widetilde{\epsilon_{i1}} = y_{i1} - \mathbf{x}_{i1}^{\mathrm{T}} \widetilde{\boldsymbol{\beta}^{OLS}}$ for every $i = 1, \ldots, n$.

4. Repeat Step 2 for the new set of residuals $\{|\widetilde{\epsilon_{11}}|, \ldots, |\widetilde{\epsilon_{n1}}|\}$.

5. Apply the lasso to $y_{i1} = \alpha_{i0} + \mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_{i1}$ but only for the n_{pure} observations. The regularization parameter for this step is the value that minimizes the extended BIC (EBIC) criterion.

6. The units for which α_{i0} was estimated to be nonzero are removed from the dataset completely.

7. Apply the panel lasso to $y_{it} = \alpha_{i0} + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{it}$ for all the remaining units i and the remaining time periods $t = 2, \ldots, T$. The regularization parameter for this step is the value that minimizes the EBIC criterion.

8. Determine the set $\widehat{I_0}$ and apply the two-step panel lasso estimator.

The first five steps can be thought of as applying the lasso to a testing set. The reasoning behind Steps 1 to 5 is to try to select the subset of the data that would most likely have $\alpha_{i0} = 0$. If the ith unit is of this type, then it is quite likely that the absolute value of the residuals in Steps 1 and 3 would be quite small relative to the other two types of incidental parameters. Steps 3 and 4 essentially repeat the first two steps to reduce the possibility of selecting units with a large absolute value for the idiosyncratic error. Notice that the first five steps use only the earliest information available in the panel. In contrast, the final three steps use all the remaining information in the panel. Finally, note that Step 5 is really a special case of the panel lasso for $T = 1$.

The extended BIC criterion, proposed by Chen and Chen (2008), modifies the usual BIC criterion so that the latter can still be applied in the context where the number of regressors P grow with sample size at some polynomial rate, i.e., $P = O\left(n^k\right)$ for some $k > 0$. The EBIC is indexed by some $\phi \in [0, 1]$ and is given by

$$BIC_{\phi}(s) = BIC(s) + 2\phi \log \begin{pmatrix} P \\ s \end{pmatrix},$$

where s is the size of the model. Notice the extra term in the criterion. This extra term penalizes models that are too large because the model space is growing large as the number of regressors also grow with sample size. Chen and Chen (2008) have shown the selection consistency of the extended BIC and when $P = O\left(n^k\right)$ and $\phi > 1 - 1/(2k)$. In the panel data case, we have $k = 1$ and we have to set $\phi > 1/2$.

The only remaining issue is that the user has to specify n_{pure}. Choosing the size of n_{pure} will ultimately depend on the user's faith in the sparsity assumption. If n is very large, n_{pure} can be set at a relatively small value. The value of n_{pure} should not be too small or too large for two reasons. First, there should still be enough degrees of freedom so that we are still able to implement Step 1, i.e., there should be enough observations so that $\boldsymbol{\beta}$ can still be estimated.[3] Second, the data-based procedure might produce a regularization parameter that overshrinks the large incidental parameters.

[3] One should take care that there are enough degrees of freedom so that the coefficients of a large but finite number of regressors can still be estimated.

3 Monte Carlo

In this section, I study the finite-sample performance of the two-step panel-lasso estimator. I start with the dynamic linear panel data model because this is often used in empirical applications. Furthermore, the empirical application discussed in Section 4 involves the estimation of a dynamic linear panel data model. The experiments follow the Monte Carlo design of Bun and Sarafidis (2015). Their design attempts to encompass existing Monte Carlo designs while ensuring comparability across different simulations. They consider the following model for $i = 1, \ldots, n$ and $t = 1, \ldots, T$:

$$\begin{aligned} y_{it} &= 0.8y_{i,t-1} + 0.2x_{it} + \alpha_i + \varepsilon_{it}, \\ x_{it} &= 0.95x_{i,t-1} + 0.25\alpha_i + \upsilon_{it}, \\ \upsilon_{it} &= \nu_{it} - 0.1\varepsilon_{it}. \end{aligned}$$

I set n to be either 50 or 1000. Assume that $\varepsilon_{it} \overset{iid}{\sim} N(0,1)$ and $\nu_{it} \overset{iid}{\sim} N\left(0, \sigma_\nu^2\right)$. To better control the experimental conditions, they suggest fixing the values of four parameters: the signal-to-noise ratio (SNR), the variance ratio (VR) and the correlation between the deviation of the initial condition of the x (and y) process(es) from its long run steady state path and the level of the steady state path itself (r_x and r_y respectively). The signal-to-noise ratio SNR represents the additional variance provided by the explained portion of the model conditional on α_{i0} after netting out the variance of ε_{it}. The variance ratio VR measures the relative magnitudes of the cumulative impact of the two error components on the average variance of y_{it} over time.[4]

I set $SNR = 3$, $VR = 100$, and $r_x = r_y = 0.5$. As a consequence, there is a lot of noise coming from α_{i0} relative to other components and the initial conditions are above the long-run steady state path. I set 40 periods for burn-in. The incidental parameters α_{i0} are iid draws from the following mixed discrete-continuous distribution:

$$\alpha_{i0} = \begin{cases} 0 & \text{with probability } p_0 \\ W_1\left(0.5 + W_2\right) & \text{with probability } p_1 \\ U\left[-0.5, 0.5\right] & \text{with probability } 1 - p_0 - p_1 \end{cases},$$

where $W_1 = -1$ with probability 0.75 and $W_1 = 1$ with probability 0.25 and W_2 are iid draws from the exponential distribution with mean κ. I choose the value of κ so that I would be able to match the standard deviation of the distribution for the incidental parameters and the distribution of ν_{it} (denoted by σ_η and σ_ν respectively). The chosen values for κ can be found in the tables containing the simulation results.

I carried out all computations with `R` and the `glmnet` package (see Friedman, Hastie,

[4] For more details on the derivation of these quantities, I refer the reader to Bun and Sarafidis (2015).

and Tibshirani (2010).[5] In the Monte Carlo experiments, I set $T = 2$ where initial condition y_{i0} is available. Since there are no moment conditions that can be used to construct a GMM estimator, I compare the two-step panel lasso estimator to the pooled OLS estimator which treats all $\alpha_{i0} = 0$ and the pooled OLS estimator where I only use the units for which $\alpha_{i0} = 0$. I expect that the pooled OLS estimator which treats all $\alpha_{i0} = 0$ to be inconsistent. The pooled OLS estimator where I only use the units for which $\alpha_{i0} = 0$ is also called the oracle estimator. Clearly, the oracle should be consistent since it uses true knowledge not available to the econometrician.

I now describe the ability of the first step of the panel lasso to predict the type of incidental parameter in Table 1. The table provides us with an understanding of the theoretical results when applied to finite samples. The reported statistics include the mean biases and standard deviation for the estimators. I also report the rejection rates for individual Wald t-tests of the true null hypotheses $\gamma = 0.8$ and $\beta = 0.2$.

I consider two designs, referred to as Designs A and B. Designs A and B consider the situation where $(p_0, p_1) = (0.966, 0.02)$ and $(p_0, p_1) = (0.83, 0.1)$, respectively. The latter design allow us to consider what happens when the conditions on the number of "bounded" and "large" incidental parameters are violated. The former design is well within the conditions specified in the theorems discussed in the previous section.

The results in the table are in line with what was developed in the Section 2. It would seem that varying n_{pure}/n did not matter so much.[6] Furthermore, the pooled OLS estimator seems to be tracking the behavior of the panel lasso estimator when $n = 50$. Once we increase the value of n twenty-fold, we see that the inference properties of the pooled OLS estimator has suffered relative to the panel lasso. Clearly, the oracle is performing the best in every aspect.

The results in Table 1 may give the impression that the proposed estimator is not performing well when there is a larger sample size. The consistency of the panel lasso estimator requires that the number of large incidental parameters is bounded as sample size grows large. There are relatively more draws for the large incidental parameters when $n = 1000$ compared to $n = 50$, especially when one looks at Design B.

4 Inequality and income growth

The relationship between inequality and income growth has long been a subject of intense economic debate. Kuznets (1955), one of the top 20 papers chosen to celebrate the centennial of the American Economic Review, precisely deals with this relationship. The first page alone already lists down the key issues with studying this relationship. I interpret the approach by van der Weide and Milanovic (2014) as one way to address some of the issues

[5] R scripts used for the computations are available upon request.

[6] Varying ϕ in the EBIC did not change the results much either.

Table 1: Finite sample performance of estimators from 1000 replications

	Mean bias γ		Mean bias β		Sd γ		Sd β		Rej. Rate $\gamma = 0.8$		Rej. rate $\beta = 0.2$	
Design A ($\kappa = 7.74$)	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$
Lasso $n_{pure}/n = 0.5, \phi = 1$	0.0170	0.0124	0.0071	0.0080	0.0795	0.0167	0.1232	0.0246	0.0590	0.1130	0.0710	0.0580
Lasso $n_{pure}/n = 0.9, \phi = 1$	0.0169	0.0124	0.0082	0.0080	0.0775	0.0167	0.1225	0.0245	0.0700	0.1140	0.0750	0.0580
Pooled OLS	0.0109	0.0124	0.0078	0.0080	0.0552	0.0122	0.0805	0.0170	0.0530	0.2020	0.0470	0.0730
Oracle	0.0003	-0.0008	0.0009	-0.0004	0.0575	0.0127	0.0838	0.0170	0.0540	0.0560	0.0490	0.0420
Design B ($\kappa = 3.367$)												
Lasso $n_{pure}/n = 0.5, \phi = 1$	0.0444	0.0463	0.0308	0.0302	0.0705	0.0142	0.1190	0.0244	0.1190	0.8990	0.0680	0.2470
Lasso $n_{pure}/n = 0.9, \phi = 1$	0.0463	0.0463	0.0291	0.0302	0.0693	0.0142	0.1189	0.0244	0.1210	0.8930	0.0700	0.2420
Pooled OLS	0.0441	0.0463	0.0300	0.0300	0.0475	0.0102	0.0790	0.0172	0.1750	0.9930	0.0630	0.4300
Oracle	-0.0014	-0.0008	0.0034	-0.0007	0.0631	0.0135	0.0902	0.0194	0.0590	0.0580	0.0540	0.0520

Note: The implied $\sigma_\eta = 1.477$ and $\sigma_\nu = 0.425$.

with studying this relationship, in particular, the need to start with the family as the unit of analysis. A casual Google search of these keywords will already point out the immense number of studies (usually at the country level) devoted to studying this relationship. As more and more data are collected and disaggregated, we see more complicated methods being applied like those that involve dynamic linear panel data methods.

In this section, I revisit the evidence found by van der Weide and Milanovic (2014) where high levels of inequality reduce the income growth of the poorest percentiles of the distribution. Instead of using readily available aggregate measures of inequality and income, they construct these aggregate measures using state-level data from the United States. Individual-level data from the Integrated Public Use Microdata Survey for 1960, 1970, 1980, 1990, 2000, and 2010 were used to construct state-level measures of income per capita (`lnyxx`), inequality (`gini`), educational shortfalls (`edushort1518`), educational attainment beyond the college level (`edu_ms_age2139`), share of women outside the labor force (`olf_female`), share of household members that are too young (`age015`) and too old to work (`age65`). Details regarding the computation and definition of these variables can be found in their paper.

van der Weide and Milanovic (2014) estimate a Solow-type growth regression at the state level that includes a measure of inequality as one of the regressors. The parameter of interest is the effect of income inequality on income growth. They estimate (1) with variables defined as follows:

1. y_{it} represents income growth at specific percentiles. This is coded as `dlnyxx` where $xx \in (05, 10, 25, 50, 75, 90, 95, 99)$.[7]

2. $\mathbf{x}_{it}^{\mathrm{T}}$ contain the first-order lags of `gini`, `edu_ms_age2139`, `edushort1518`, `age015`, `age65`, `olf_female`, `lnyxx`, and time dummies. They also use two alternative measures of `gini`, namely, the state-level Gini of the bottom 40% (`gini_b40`) and top 40% (`gini_t40`) of the population.[8]

3. They consider data from $n = 51$ states and $T = 5$ time periods (representing every decade since 1960). Alaska and the District of Columbia were considered outliers and were excluded from the sample.

I apply the panel lasso estimator to the model just described for $T = 2$ and for all states.[9] Why would the panel lasso be appropriate in this empirical application? The model chosen in the empirical application can be thought of as a growth regression possibly based on an augmented Solow growth model. Just as I discussed in Section 2 where I introduced the

[7] It is unclear whether estimating separate regressions for each percentile is preferable over quantile regressions. I leave this to future work.

[8] They wanted to "unpack" the effect of inequality at the bottom and at the top on income growth at different percentiles.

[9] Other configurations were implemented but the patterns obtained in Figures 1, 2, and **??** remain.

panel lasso, we can think of the incidental parameters as representing the suitability or fit of the Solow growth model to the data. In particular, large non-zero incidental parameters represent states for which the growth regression may not be a good approximation. The bounded incidental parameters represent moderate state-specific deviations from the growth model that can be shrunk toward zero (as this would have no effect on the estimator properties, ast least asymptotically). An alternative justification is that sparsity can be useful when n and T are both small. The Monte Carlo experiments already provide some evidence in this regard. Furthermore, the two-step panel lasso estimator can be applied even for $T = 2$ and even accommodates contemporaneously exogenous regressors. For the moment, there is no estimator that could match these advantages.

The results of the one-step panel lasso estimator indicate that it is possible to pool all the states, regardless of whether I use `gini` as the inequality measure or `gini_b40` and `gini_t40` as the inequality measures. As a result, the exclusion of Alaska and DC from the sample by van der Weide and Milanovic (2014) may be unwarranted given the results of the one-step panel lasso estimator. The overall conclusion seems to be that heterogeneity across states may not be as large as one might think.

I then estimate using the proposed two-step panel lasso estimator for $T = 2$. Robust standard errors are used to construct the confidence intervals. Since the main interest of van der Weide and Milanovic (2014) is the effect of inequality on income growth, I only present 95% confidence intervals for the slopes of `gini`, `gini_b40`, and `gini_t40`. I set $n_{pure}/n = 0.8$ and $\phi = 1$ for the computation of the regularization parameter. I also report results from the pooled OLS estimator where there is only an overall intercept and no state-specific fixed effects. All other results are available upon request.

Figure 1 already gives an impression that the effect of inequality on income growth for the top 50% of the population has not changed so much over time. Although the effect of inequality is mostly positive for the top 50% of the population, the estimated effects are not as large as suggested by the system GMM results of van der Weide and Milanovic (2014). In contrast, the effect of inequality on income growth for the bottom 50% of the population has had substantial changes over time. If we compare data from 1970-1980 and the decades after, we see that even if the estimated effects are negative (and sometimes close to zero when looking at the median), the absolute values of these estimated effects are getting smaller over time.

Results of pooled OLS estimation can be found in Figure 2. The results are substantially different from Figure 1 in two respects. First, the confidence intervals obtained by pooled OLS for 1970-1980 are strikingly different from those obtained from the two-step panel lasso. Second, the standard errors are much larger for the panel lasso. I interpret the results of Figure 2 as evidence that we may have to conduct a separate analysis of the 1970-1980 decade. Furthermore, the figure casts doubt on whether there is parameter constancy over time. The panel lasso has somehow stabilized this parameter inconstancy.

Whichever figure one uses, there seems to be a sharp change in the relationship of inequality (whether bottom or top or as a whole) and income growth across all percentiles after 1970-1980. After this sharp change, this relationship has not changed so much after 1990, especially at the top percentiles. Most of the estimated effects of bottom inequality on income growth are statistically different from zero, especially for the percentiles above the median in recent years. I find that higher bottom inequality has a positive relationship with income growth at the top percentiles just like van der Weide and Milanovic (2014) but the magnitudes are slightly smaller. The estimated effects of top inequality on income growth are statistically not different from zero, especially for the percentiles above the median. If we look at Figures 1 and 2, there is reason to be optimistic because of the gradual reduction in the absolute effect of inequality (whether bottom or top or as a whole) on income growth.

To summarize, the results are strikingly different from the reported impression of massive inequality during 1990-2010. The absolute effect of bottom or top inequality on income growth has been getting smaller across time and across percentiles, especially when one looks at the bottom 50% of the population. The sharp change after the 1970-1980 decade might be driving the rather negative results (in the sense that they find that inequality is good for the rich but not for the poor) of van der Weide and Milanovic (2014). The notion that inequality (whether bottom or top) benefits only the rich may be a lot more nuanced than we think.

5 Concluding remarks

We show how the penalized least squares approach in the presence of incidental parameters of Fan, Tang, and Shi (2012) can be extended to panel data models. Not all of their results survive the extension. The most serious change in terms of consistency and valid inference is the need to bound the number of "large" incidental parameters by a constant. Despite this, I was able to allow for contemporaneously exogenous regressors. The sparsity of incidental parameters has been useful in deriving consistent estimators for the structural parameters. They come at the cost of specifying a particular structure for the asymptotic growth in the different types of incidental parameters in order to obtain consistency and asymptotic normality. The latter has been problematic in the context of estimators that encourage sparsity, as discussed extensively by Leeb and Pötscher (2005; 2008).

I also propose a data-based procedure for choosing the regularization parameter which uses the extended BIC criterion since the usual BIC criterion is inconsistent when the number of parameters grow at a polynomial rate with sample size. Results of Monte Carlo experiments indicate good finite sample performance of the two-step panel lasso estimator for very small T. Unfortunately, departures from the assumed sparsity of the incidental parameters create substantial problems for consistent estimation and valid inference. As a result, the two-step panel lasso estimator is unable to match the performance of the oracle

estimator but is preferable to simply using pooled OLS.

I also use the two-step panel lasso estimator to shed light on the relationship between inequality and income growth by revisiting the evidence of van der Weide and Milanovic (2014). The small sample size deters us from making stronger conclusions about what makes the excluded states different from the others. Perhaps the most optimistic aspect of the results is the gradual move toward the reduced impact of inequality on income growth across all percentiles.

Although the focus has been on fixed-T consistent estimation, an analysis of the performance of the penalized least squares estimator under alternative asymptotic embeddings such as letting $n, T \to \infty$ jointly or at a particular rate, say $n/T \to c \in (0, \infty)$ would be of practical value. The effect of a larger value of T might help us reduce the restrictions on the growth in the different types of incidental parameters. This analysis will also give insight as to the statistical benefits of a repeated observation and determine whether the cost of collecting panel data is justifiable. Furthermore, the derivation of the asymptotic properties of the two-step panel lasso estimator uses results from seemingly unrelated regressions, as introduced by Zellner (1962). Seemingly unrelated regressions give a natural framework for allowing varying coefficients not just for the intercept term but for the slope coefficients as well. By allowing for this extension, we may be able to develop alternative estimators for the varying coefficients model. Extensions to the case of nonlinear panel data models will also be needed. Finally, linking the properties of the pretest estimator after a test of poolability to that of the two-step panel lasso estimator may also be of practical value. I leave all these to future research.

References

Arellano, M and S Bonhomme (2011). Nonlinear panel data analysis. *Annual Review of Economics* **3**, 395–424.

Bonhomme, S (2012). Functional Differencing. *Econometrica* **80**.(4), 1337–1385.

Bonhomme, S and E Manresa (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica* **83**.(3), 1147–1184.

Browning, M and JM Carro (2010). Heterogeneity in dynamic discrete choice models. *Econometrics Journal* **13**.(1), 1–39.

Bun, MJG and V Sarafidis (2015). "Chapter 3 – Dynamic Panel Data Models". In: *The Oxford Handbook of Panel Data*. Ed. by BH Baltagi. Oxford University Press, pp.76–110.

Chen, J, J Gao, and D Li (2013). Estimation in Single-Index Panel Data Models with Heterogeneous Link Functions. *Econometric Reviews* **32**.(8), 928–955.

Chen, J and Z Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**.(3), 759–771.

Fan, J, R Tang, and X Shi (2012). Partial Consistency with Sparse Incidental Parameters. *ArXiv e-prints*.

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

Kock, AB (2013). Oracle Efficient Variable Selection in Random and Fixed Effect Panel Data Models. *Econometric Theory* **29** (01), 115–152.

Kock, AB (2014). *Oracle inequalities and Variable Selection in High-Dimensional Panel Data Models*. Tech. rep. `https://sites.google.com/site/andersbkock/LassoPanel.pdf`.

Kock, AB and H Tang (2014). *Inference in high-dimensional dynamic panel data models*. Tech. rep. `https://sites.google.com/site/andersbkock/KockTang_v11_20141224_3.pdf`.

Kuznets, S (1955). Economic growth and income inequality. *The American Economic Review* **45**.(1), 1–28.

Leeb, H and BM Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59.

Leeb, H and BM Pötscher (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142**, 201–221.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. `http://www.R-project.org`.

Sarafidis, V and N Weber (2011). *A Partially Heterogeneous Framework for Analyzing Panel Data*. Tech. rep.

van der Weide, R and B Milanovic (2014). *Inequality is bad for growth of the poor (but not for that of the rich)*. Policy Research Working Paper Series. The World Bank. `http://ideas.repec.org/p/wbk/wbrwps/6963.html`.

Wooldridge, JM (2010). *Econometric Analysis of Cross-Section and Panel Data*. MIT Press.

Zellner, A (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.

A Proofs of some of the main results

A.1 Proof of Lemma 1

The argument follows Fan, Tang, and Shi (2012) but with some modifications and corrections. Let C be an arbitrary (small) positive number and $\boldsymbol{\beta} \in B_C\left(\boldsymbol{\beta}_0\right)$.

We first prove that $S_{10} = S_{10}^*, S_{20} = \varnothing, S_{30} = \varnothing$ wpg 1. It is always true that $S_{10} \subseteq S_{10}^*$. Thus we have to show that $\Pr\left(S_{10} \supseteq S_{10}^*\right) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1\leq i\leq n} \|\bar{\mathbf{x}}_i\|_2 \leq \kappa_n\right\}$ and $\mathscr{D} = \{s+1 \leq i \leq n : |\bar{\epsilon}_i| < \gamma_n\}$. Note that $\Pr\left(S_{10} \supseteq S_{10}^*\right) \geq \Pr\left(S_{10} \supseteq S_{10}^* \mid \mathscr{B}\right)\Pr(\mathscr{B})$.

Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr\left(S_{10} \supseteq S_{10}^* \mid \mathscr{B}\right) \to 1$. For large n, $\Pr\left(S_{10}^* \subseteq \mathscr{D}\right) \to 1$. Conditional on $\mathscr{B}$, $|\bar{\epsilon}_i| \leq \gamma_n$ implies that

$$\left|\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) + \bar{\epsilon}_i\right| \leq \left|\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right)\right| + |\bar{\epsilon}_i| \leq \left\|\bar{\mathbf{x}}_i^{\mathrm{T}}\right\|_2 \left\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right\|_2 + |\bar{\epsilon}_i| \leq \kappa_n \sqrt{d} C + \gamma_n,$$

and we have $\kappa_n \sqrt{d} C + \gamma_n \leq \lambda$ for large n by (13). As a result, $\mathscr{D} \subseteq S_{10}$. Thus, $\Pr\left(S_{10}^* \subseteq \mathscr{D} \subseteq S_{10} \mid \mathscr{B}\right) \to 1$. Since $S_{10} \cup S_{20} \cup S_{30} = S_{10}^*$ is always true, we must have $S_{20} = \varnothing$ and $S_{30} = \varnothing$ wpg 1.

Next we prove that $S_{11} = \varnothing, S_{21} = S_{21}^*, S_{31} = S_{31}^*$ wpg 1. We show that $S_{21} = S_{21}^*$ wpg 1 as the case for $S_{31} = S_{31}^*$ wpg 1 is analogous. . Let $S_{211} = S_{21} \cap S_{21}^*$ and $S_{212} = S_{21} \cap \left(S_{21}^*\right)^c$. It suffices to show that $\Pr\left(S_{211} = S_{21}^*\right) \to 1$ and $\Pr\left(S_{212} = \varnothing\right) \to 1$. It is always true that $S_{21} \subseteq S_{21}^*$. Thus we have to show that $\Pr\left(S_{21} \supseteq S_{21}^*\right) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1 \leq i \leq n} \|\bar{\mathbf{x}}_i\|_2 \leq \kappa_n\right\}$ and $\mathscr{D} = \{1 \leq i \leq s_1 : \bar{\epsilon}_i \geq -\gamma_n\}$. Note that $\Pr\left(S_{21} \supseteq S_{21}^*\right) \geq \Pr\left(S_{21} \supseteq S_{21}^* \mid \mathscr{B}\right)\Pr(\mathscr{B})$. Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr\left(S_{211} \supseteq S_{21}^* \mid \mathscr{B}\right) \to 1$. For large n, $\Pr\left(S_{21}^* \subseteq \mathscr{D}\right) \to 1$. Conditional on $\mathscr{B}$ and noting that $\alpha_{i0} > 0$, $\bar{\epsilon}_i > -\gamma_n$ implies that

$$\alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) + \bar{\epsilon}_i > \alpha^* - \kappa_n \sqrt{d} C - \gamma_n,$$

and we have $\alpha^* - \kappa_n \sqrt{d} C - \gamma_n \geq \lambda$ for large n by (13). As a result, $\mathscr{D} \subseteq S_{211}$. Thus, $\Pr\left(S_{21}^* \subseteq \mathscr{D} \subseteq S_{211} \mid \mathscr{B}\right) \to 1$.

Now, we show that $\Pr\left(S_{212} = \varnothing\right) \to 1$. It is always true that $\varnothing \subseteq S_{212}$. Thus we have to show that $\Pr\left(S_{212} \subseteq \varnothing\right) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1 \leq i \leq n} \|\bar{\mathbf{x}}_i\|_2 \leq \kappa_n\right\}$ and $\mathscr{D} = \{1 \leq i \leq s_1 : \bar{\epsilon}_i > \gamma_n\}$. Note that $\Pr\left(S_{212} \subseteq \varnothing\right) \geq \Pr\left(S_{212} \subseteq \varnothing \mid \mathscr{B}\right)\Pr(\mathscr{B})$. Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr\left(S_{212} \subseteq \varnothing \mid \mathscr{B}\right) \to 1$. For large n, $\Pr\left(\mathscr{D} \subseteq \varnothing\right) \to 1$. Conditional on $\mathscr{B}$, noting that $\alpha_{i0} < 0$ and $\gamma_n - \alpha^* + \kappa_n \sqrt{d} C < \lambda$ for large n by (13), $\alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right) + \bar{\epsilon}_i > \lambda$ implies that

$$\bar{\epsilon}_i > \lambda - \alpha_{i0} - \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) > \lambda + \alpha^* - \kappa_n \sqrt{d} C > \gamma_n.$$

As a result, $S_{212} \subseteq \mathscr{D}$. Thus, $\Pr\left(S_{212} \subseteq \mathscr{D} \subseteq \varnothing \mid \mathscr{B}\right) \to 1$. Along with $\Pr\left(S_{21}^* \subseteq S_{211} \mid \mathscr{B}\right) \to 1$, we have $S_{21} = S_{21}^*$ wpg 1.

Finally, we prove that $S_{12} = S_{12}^*, S_{22} = \varnothing, S_{32} = \varnothing$ wpg 1. It is always true that $S_{12} \subseteq S_{12}^*$. Thus we have to show that $\Pr\left(S_{12} \supseteq S_{12}^*\right) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1 \leq i \leq n} \|\bar{\mathbf{x}}_i\|_2 \leq \kappa_n\right\}$ and $\mathscr{D} = \{s_1 + 1 \leq i \leq s : |\bar{\epsilon}_i| < \gamma_n\}$. Note that $\Pr\left(S_{12} \supseteq S_{12}^*\right) \geq \Pr\left(S_{12} \supseteq S_{12}^* \mid \mathscr{B}\right)\Pr(\mathscr{B})$. Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr\left(S_{12} \supseteq S_{12}^* \mid \mathscr{B}\right) \to 1$. For large n, $\Pr\left(S_{12}^* \subseteq \mathscr{D}\right) \to 1$. Conditional on $\mathscr{B}$, $|\bar{\epsilon}_i| \leq \gamma_n$ implies that

$$\alpha_{i0} - \kappa_n \sqrt{d} C - \gamma_n \leq \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) + \bar{\epsilon}_i \leq \alpha_{i0} + \kappa_n \sqrt{d} C + \gamma_n,$$

and we have $-\lambda - \alpha_{i0} + \kappa_n \sqrt{d} C < -\gamma_n$ and $\lambda - \alpha_{i0} - \kappa_n \sqrt{d} C > \gamma_n$ for large n by (13). As a

result, $\mathscr{D} \subseteq S_{12}$. Thus, $\Pr\left(S_{12}^* \subseteq \mathscr{D} \subseteq S_{12} \mid \mathscr{B}\right) \to 1$. Since $S_{12} \cup S_{22} \cup S_{32} = S_{12}^*$ is always true, we must have $S_{22} = \varnothing$ and $S_{32} = \varnothing$ wpg 1.

A.2 Proof of Theorem 1

We now analyze every term in (11) after substituting in (12) and applying Lemma 1:

1. Collect all the terms that involve α_{i0}. Under A3 and A4, we have

$$\left\| \frac{1}{n} \sum_{i \in S_{12}^*} \bar{\mathbf{x}}_i \alpha_{i0} \right\|_2 \leq \frac{1}{n} \sum_{i \in S_{12}^*} \|\bar{\mathbf{x}}_i\|_2 |\alpha_{i0}| \leq \frac{s - s_1}{n} \kappa_n \gamma_n.$$

   Provided that $s - s_1 = o\left(n / (\kappa_n \gamma_n)\right)$, we have $\frac{1}{n} \sum_{i=1}^{n} \bar{\mathbf{x}}_i \alpha_{i0} = o_p(1)$.

2. By the law of large numbers along with A1,

$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{x}_{it} \epsilon_{it} \xrightarrow{p} \lim_{n \to \infty} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}(\mathbf{x}_{it} \epsilon_{it}) = \mathbf{0}.$$

   The latter equality follows from A2-1 or A2-2 or even strict exogeneity.

3. Strict exogeneity of $\mathbf{x}_{it}$ allows us to conclude that $\frac{1}{n} \sum_{i=1}^{n} \bar{\mathbf{x}}_i \bar{\epsilon}_i \xrightarrow{p} \mathbf{0}$. If any of the variables in $\mathbf{x}_{it}$ is predetermined or weakly exogenous, the argument has to change slightly, i.e.

$$\left\| \frac{1}{n} \sum_{i \in S_{21}^* \cup S_{31}^*} \bar{\mathbf{x}}_i \bar{\epsilon}_i \right\|_2 \leq \frac{1}{n} \sum_{i \in S_{21}^* \cup S_{31}^*} \|\bar{\mathbf{x}}_i \bar{\epsilon}_i\|_2 \leq \frac{s_1}{n} \kappa_n \gamma_n.$$

   The latter is $o(1)$ when $s_1 = o\left(n / (\kappa_n \gamma_n)\right)$.

4. Let $S = S_{21}^*, S_{31}^*$. Note that

$$\left\| \frac{\lambda}{n} \sum_{i \in S} \bar{\mathbf{x}}_i \right\|_2 \leq \frac{\lambda}{n} \sum_{i \in S} \|\bar{\mathbf{x}}_i\|_2 \leq \frac{\lambda}{n} s_1 \kappa_n = \frac{\lambda}{\sqrt{n}} s_1 \frac{\kappa_n}{\sqrt{n}}.$$

   The latter will only be $o_p(1)$ when A3 holds, λ obeys (13), and $s_1 = O(1)$.

A.3 Proof of Lemma 2

Let $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2$. Assume that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. Define the following probabilities:

$$T_0 \quad = \quad \Pr\left(\bigcap_{i=s+1}^{n} \left\{ \left| \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i \right| \leq \lambda \right\} \right)$$

$$
\begin{aligned}
T_1 &= \Pr\left(\bigcap_{i=1}^{s_1}\left\{\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\bar{\epsilon}_i>\lambda\right\}\right)\\
T_2 &= \Pr\left(\bigcap_{i=s_1+1}^{s}\left\{\left|\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\bar{\epsilon}_i\right|\leq\lambda\right\}\right)
\end{aligned}
$$

Notice that $\Pr(\mathscr{E})=T_0T_1T_2$. Therefore, to show that $\Pr(\mathscr{E})\to 1$, it suffices to show that $T_0\to 1$, $T_1\to 1$, and $T_2\to 1$. Note that

$$
\begin{aligned}
1-T_1 &= \Pr\left(\bigcup_{i=1}^{s_1}\left\{\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\bar{\epsilon}_i\leq\lambda\right\}\right)\\
&\leq \underbrace{\Pr\left(\bigcup_{i\in S_{21}^*}\left\{\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\bar{\epsilon}_i\leq\lambda\right\}\right)}_{T_{11}}+\underbrace{\Pr\left(\bigcup_{i\in S_{21}^*}\left\{\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\bar{\epsilon}_i\leq\lambda\right\}\right)}_{T_{12}}.
\end{aligned}
$$

We just have to show that $T_{11}\to 0$ and $T_{12}\to 0$. Define $\mathscr{C}=\left\{\left\|\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right\|_2<\overline{C}\right\}$ where $\overline{C}=(\alpha-1)/\left(M\alpha\sqrt{d}\right)>0$, for some choice of M. Note that

$$
\begin{aligned}
T_{11} &\leq \Pr\left(\bigcup_{i\in S_{21}^*}\left[\left\{\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\bar{\epsilon}_i\leq\lambda\right\}\cap\mathscr{C}\right]\right)+\Pr(\mathscr{C}^c)\\
&\leq \Pr\left(\bigcup_{i\in S_{21}^*}\left[\left\{\bar{\epsilon}_i\leq\lambda-\alpha^*+\kappa_n\overline{C}\sqrt{d}\right\}\right]\right)+\Pr(\mathscr{C}^c)\\
&\leq \Pr\left(\bigcup_{i\in S_{21}^*}\left[\{\bar{\epsilon}_i\leq-\gamma_n\}\right]\right)+\Pr(\mathscr{C}^c)\\
&\leq s_1\Pr(\{\bar{\epsilon}_i\leq-\gamma_n\})+\Pr(\mathscr{C}^c)\to 0.
\end{aligned}
$$

The first inequality follows from the law of total and probability and the monotonicity of the probability function. The second inequality follows from the definition of $\mathscr{C}$ and the characteristics of the incidental parameters belong to the set S_{21}^*. The third and fourth inequalities follows from the specification of the regularization parameter found in (13) and subadditivity. The convergence to zero follows from assumption A3 and the consistency of the panel lasso. An analogous derivation will show that $T_{12}\to 0$.

To show that $T_0\to 1$, note that

$$
\begin{aligned}
T_0 &\geq \Pr\left(\bigcap_{i=s+1}^{n}\left\{-\lambda-\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)\leq\bar{\epsilon}_i\leq\lambda-\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)\right\}\cap\mathscr{C}\right)\\
&\geq \Pr\left(\bigcap_{i=s+1}^{n}\left\{-\gamma_n\leq\bar{\epsilon}_i\leq\gamma_n\right\}\right)\to 1.
\end{aligned}
$$

The first inequality follows from the monotonicity of the probability function and some

algebra. The second inequality arises because λ obeys (13) and

$$\begin{aligned} -\lambda - \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\right) &\leq -\lambda + \left\|\bar{\mathbf{x}}_i^{\mathrm{T}}\right\|_2 \left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\right\|_2 \leq -\lambda + \kappa_n \overline{C}\sqrt{d} \\ &\leq -\lambda + \lambda M \overline{C}\sqrt{d} = \lambda\left(M\overline{C}\sqrt{d} - 1\right) < \lambda\left(-\frac{1}{\alpha}\right) \leq -\gamma_n \end{aligned}$$

for the choice of $\overline{C}$ indicated earlier.

Figure 1: 95% confidence intervals obtained from the panel lasso

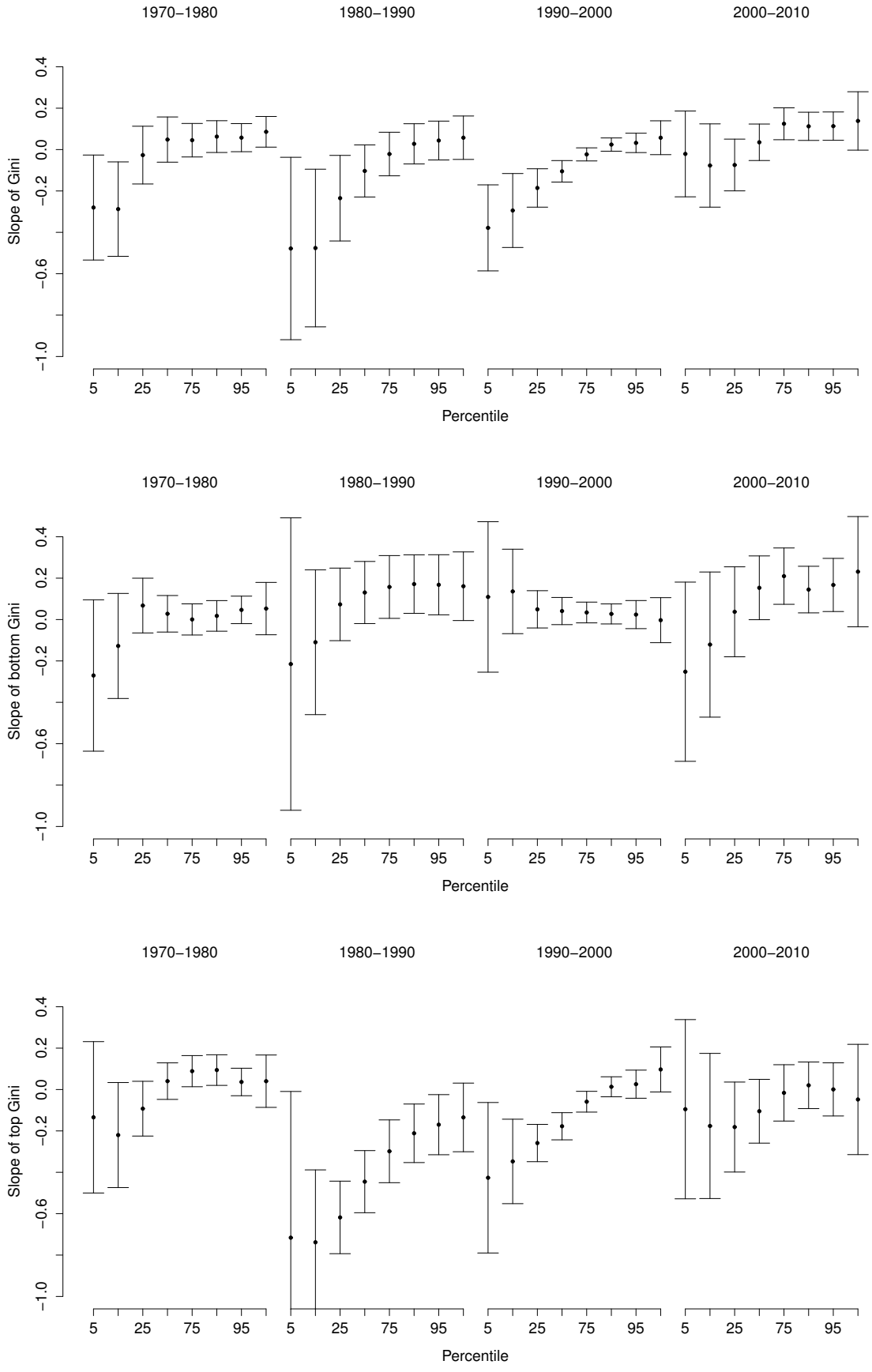

Figure 2: 95% confidence intervals obtained from pooled OLS