Advanced Microeconometrics

Prof. D. Margolis

High-dimensional panel data using the Lasso

Jérémy L'Hour*, Nicolas Saleille†
ENSAE ParisTech and Paris School of Economics

January 7, 2015

Contents

*jeremy.l.hour@ensae-paristech.fr
†nicolas.saleille@ensae-paristech.fr

1 Introduction

The use of **panel data** is extremely appealing in microeconometrics because they allow to estimate the impact of a variable of interest while accounting for individual-specific unobserved heterogeneity and disentangling components of variance without making assumptions that are too restrictive (Arellano, 2003). For example, one of the most commonly used method is the linear fixed effect model that allows a part of the error term to be correlated with the time-varying regressors. In this case, no assumption other than linearity is made regarding the individual-specific heterogeneity which strengthens the robustness of the results compared to the random effect model for example.

Panel data are also likely to be high-dimensional data in the sense that they may have a large number of time-varying regressors. Typically, there may be multiple levels of observation that may or may not be nested: the individual, the county, the state, the country, the firm, etc. Examples of such studies include Abowd *et al.* (1999). But even with a small dataset, one can be quickly faced with a high-dimensional problem, when several transformations of a variable or interactions between variables are considered to allow for flexible effects to be captured, *e.g.* in consumption models where the income is interacted with a social category dummy, or if sieve approximations are used to estimate a non-parametric model. We will clarify later the contexts within which the number of regressors may be larger, or at least proportional to the number of observations.

Estimating high-dimensional models is not straightforward because for example ordinary least squares give a perfect fit when there are as many explanatory variables as observations in the dataset. To deal with the variable selection problem in a rigorous framework and overcome the shortcomings of usual methods like OLS, several methods have been developed. Let us first clarify the problem in a simple case. Consider the following linear regression model with the usual assumptions:

$$y_i = x_i^T \beta + \epsilon_i, \forall i = 1, .., n \tag{1}$$

where y_i is the observed dependent variable for individual i, x_i is the (column) vector of p observed random regressors, β is the $p \times 1$ deterministic vector of coefficients and ϵ_i is the residual. For convenience, we also define the $n \times p$ matrix $X := \left(x_1^T, ..., x_n^T\right)$. We assume (y_i, x_i) to be iid and $\mathbb{E}(\epsilon_i | x_i) = 0$. The high-dimensional case arises when $n < p$, in which case the OLS estimator cannot be computed. For the applied econometrician, this problem may arise even if we have $p \leq n$ but p relatively large compared to n because the inverse of $X^T X$ may be ill-behaved. A convenient way to deal with high-dimensional models is to assume *approximate sparsity* in β, *i.e.* only $s << n$ elements of β are non-zero and they are sufficient to grasp the main features of the model. We call *sparsity pattern* the set of non-zero components of β: $J(\beta) := \{j : \beta_j \neq 0\}$.

Fairly recently, a lot of attention has been devoted to estimators that penalize non-zero elements of β in a way that makes computation relatively efficient, namely with a convex function. The most famous of these estimators is the Lasso of Tibshirani (1996) that uses the closest convex function - the ℓ_1-norm. The Lasso is defined as the result of the following

minimization program:

$$\min_{\beta} n^{-1} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda ||\beta||_{\ell_1} \tag{2}$$

The Lasso minimizes the residual sum of squares, exactly as in the OLS, but there is also a term that penalizes the "size" of β in order to shrink the coefficients. The tuning parameter λ sets the trade-off between a reduced bias (the OLS estimator is obtained when $\lambda = 0$) and a reduced variance (when λ is large). The Lasso has been shown to have good estimation properties with i.i.d. data, have been extended to the non-parametric and non-Gaussian cases and the optimal choice for the penalty level λ has been studied in Bickel *et al.* (2009); Meinshausen and Yu (2009); Lounici (2008); Belloni and Chernozhukov (2010); Zhao and Yu (2006); Zhang and Huang (2008).

Here, we are interested in the **application of high-dimensional methods to the particular case of panel data models**. The main references for panel data models in the small-dimensional setting are Wooldridge (2001); Arellano (2003); Magnac (2001). In these models, the setting is the following: we observe n individuals (each indexed by i) at T distinct points in time (each indexed by t), and we have the linear regression model:

$$y_{i,t} = x_{i,t}^T \beta + c_i + \epsilon_{i,t} \tag{3}$$

with $x_{i,t}$ a potentially very large vector of explanatory variables. The random variable c_i can be interpreted as the unobserved idiosyncratic characteristics of individual i that also affect y_i. In a wage equation, one can think of c_i as the ability, intelligence or social skills of the worker; in the Solow-Swan test example, c_i could be the quality of institutions as their usual measures are imperfect and one can chose not to use them; in a price equation, c_i could be non-measurable or qualitative characteristics of the good, such as the quality of the neighborhood for a flat. Note that there are n different values of c_i which means that if we have to estimate them by standard methods, we already consume n degrees of freedom. Assumptions about c_i and $\epsilon_{i,t}$ are key when considering the consistency of a given estimator of β. Let us focus on the case where the strong exogeneity assumption holds: $\mathbb{E}\left(\epsilon_{i,t}|x_{i,1}, ..., x_{i,T}, c_i\right) = 0$. The case where the random effect assumption also holds, *i.e.* $\mathbb{E}\left(c_i|x_{i,1}, ..., x_{i,T}\right) = 0$, isn't the most interesting one because the model boils down to the usual (potentially high-dimensional) pooled linear regression problem, albeit the error terms cannot be assumed to be heteroscedastic anymore. Here, we are particularly interested in the case where the random effect assumption does not hold, *i.e.* when the unobserved heterogeneity causes indeed a problem by violating the usual exogeneity assumption. This is the case for which panel data are essential to obtain consistent estimators of β, the parameter of interest.

Immediate application of the Lasso to panel data model is not immediate for at least two reasons. The first one is that independence between observations cannot be assumed any longer. Indeed, assuming that individuals behave independently from each other is not too far-fetched as it is usually done in the simplest econometric models. But observations surely are not independent over time within the same cross-sectional unit of observation, hence the need to introduce the individual-specific heterogeneity effect c in the regression equation. Falling to account for this correlation may lead to biased estimations. The second reason is that ap-

proximate sparsity may not be the best assumption to deal with fixed effects. Indeed, it would amount to assume that the individual-specific heterogeneity differs from a constant level only for a small number of individuals while for all the others this unobserved heterogeneity may be ignored. Consequently, Lasso-type methods have to be adapted to be of use in this framework. Explicit contributions to the high-dimensional panel data literature have been scarce so far. We can mention Kock (2013a,b) that assume approximate sparsity in the unobserved heterogeneity and very recently, Belloni *et al.* (2014) which is a more convincing contribution.

In section 2, we motivate and illustrate the use of high-dimensional methods in the context of panel data models and more widely inn microeconometrics. Section 3 review two main contributions to the econometrics of high-dimensional panel models. Section 4 illustrates the use of the Cluster-Lasso of Belloni *et al.* (2014) with a Monte-Carlo experiment and an application to wage data as we take another look at the union wage premium estimated in Vella and Verbeek (1998).

2 Sparsity in fixed effect panel models: economic examples

In this section, we review microeconomic examples where high-dimensional techniques could be useful. These examples are not necessarily taken from the panel data literature but we can easily imagine to have panel data for the problems that we mention.

1. **Non-parametric models:** one may want to capture a richer behaviour between y_i and x_i than the one assumed by the linear regression model. Say the true regression function is unknown and we want to estimate it: $y_i = f(x_i) + \epsilon_i$ with $\mathbb{E}(\epsilon_i|x_i) = 0$. We can approximate f by f_β a linear combination of polynomial or spline transformations of x_i that will entail a large number of regressors if we chose a high order of the polynomial transformation. In this case, there may not be sparsity in the true model, but approximate sparsity could be assumed for convenience in order to compute an estimator $\hat{f}_\beta$ of f_β by using for example a Lasso estimator. Such considerations are useful in the case of estimating returns to education where it is hard to believe in a linear relationship between earnings and years of schooling. Instead, one could more realistically assume that earnings change abruptly for those who are extremely well-educated (*e.g.* for MBA graduates) as illustrated by Belloni and Chernozhukov (2009). Another example that deals with panel data is the empirical application given in Belloni *et al.* (2014). They take another look at the effect of gun control on homicide data from Cook and Ludwig (2004) by allowing for a high-dimensional but sparse specification. They corroborate the original results and give them more robustness.

2. **"Agnostic approach" to model selection:** considering a large number of potentially significant regressors but assuming that only a few of them are actually significantly related to the outcome (*i.e.* assuming sparsity) is a way to have an agnostic look at the data. Indeed, it is often the role of economic theory to justify the use of a regressor, but the researcher may want to try all the variables available instead of relying on prior beliefs given by the economic theory (see for example Sala-i Martin (1997) in the context of finding growth determinants). Considering a high-dimensional x_i is a way to take a step back and recover the sparsity pattern instead of simply assuming it. The consistency of the Lasso in terms of recovering the true sparsity pattern has been studied in Zhao and Yu (2006); Bickel *et al.* (2009); Lounici (2008). The problem of interest in Sala-i Martin (1997) has a panel data structure since growth is observed for several countries over many years.

3. **Multiple sources of observed heterogeneity:** a high-dimensional model can be useful when one suspects multiple sources of heterogeneity but cannot pinpoint exactly which variables are to be used as controls. The most prominent example of this case is the test of the convergence hypothesis in the Solow-Swan model, *i.e.* poorer countries should experience higher growth rates as they are catching up with the richer ones. But because countries are different in many respects that can also interfere with the GDP growth rate, one cannot expect the Solow-Swan hypothesis to be verified unconditionally. This example is also illustrated in Belloni and Chernozhukov (2009). Belloni *et al.* (2013) consider a partially linear model where there is only one coefficient of interest (in their example: the effect of abortion on the crime rate) but potentially many confounding factors that have an impact both on the outcome and on the variable of interest (we call it the "policy variable"

for convenience). The interesting feature of the paper is that the confounding factors enter in a non-parametric way in the two equations which are estimated by a linear combination of many transformations of a set of variables. The key assumption is that only a small subset of variables is needed to control for the confounding factors. Then, a Lasso-type approach is used to perform model selection and Post-Lasso is used for estimation.

4. **Multiple sources of unobserved heterogeneity:** in the few cases where there is endogeneity and the econometrician has many instruments to select from, the Lasso can provide a way to better perform variable selection and improve the preciseness of the estimation. A classical example is the estimation of returns to schooling in the Angrist and Krueger (1991) dataset which becomes a high-dimensional problems when many interactions and transformations are considered ($p = 1530$!). This example is also illustrated in Belloni and Chernozhukov (2009).

3 Overcoming the curse of dimensionality via Lasso-type estimators

We now review two solutions that have been proposed in the literature.

3.1 Kock (2013a)'s estimator

3.1.1 A statistical argument

Before diving into the underlying economic assumptions that lie behind using a ℓ_1-penalized Least-Squares procedure, let us state an estimation result that would justify a bit more the idea of penalized estimators in the context of panel models. This result can be found for example in Koenker (2004). Firstly, we rewrite model 3 in its matrix form:

$$y = X\beta + Zc + \epsilon \tag{4}$$

Where Z represents a matrix full of 1s and 0s so as to identify individual i's effect in the model. Now, assume ϵ to be conditionnally distributed as $\mathcal{N}(0_{nT}, R)$ and c as $\mathcal{N}(0_{nT}, Q)$, independently from ϵ. We are in the random effect setting but it would not be very complicated to extend this to the fixed effect setting. The whole residual u, is distributed as $\mathcal{N}\left(0_{nT}, R + ZQZ^T\right)$. The GLS estimator of β is given by:

$$\hat{\beta}^{GLS} = \left(X^T\left(R + ZQZ^T\right)^{-1}X\right)^{-1}X^T\left(R + ZQZ^T\right)^{-1}y \tag{5}$$

Note that such an estimator is also the solution of the following program:

$$\min_{\beta,c} \|y - X\beta - Zc\|^2_{R^{-1}} + \|c\|^2_{Q^{-1}} \tag{6}$$

Proof of this result can be found in Koenker (2004). We can see that in this case, the optimal estimator of β is the result of a penalized Least-Squares program that shrinks individual effects c towards zero. We can notice that this penalty depends on the assumed distribution of c which is a very particular case. Starting from this observation, we could as well define another penalty reflecting a different prior belief about the distribution of c and achieve a better estimate of β even though the dimension of β is not necessarily large[1]. This is the basic idea of Kock (2013a).

[1]For the link between prior distribution in the Bayesian framework and penalized regressions, see for example

3.1.2 Kock (2013a)'s estimator

In the fixed effect and regular exogeneity setting for model 3, Kock (2013a) assumes that $c = (c_1, ..., c_n)$ is sparse and derives the properties of the so-called "Panel-Lasso" estimator for which the slope β and the individual effects c are penalized separately:

$$\min_{\beta,c} \sum_{i=1}^{n}\sum_{t=1}^{T} \left(y_{i,t} - x_{i,t}^T\beta - c_i\right)^2 + 2\lambda_{N,T}||\beta||_{\ell_1} + 2\mu_{N,T}||c||_{\ell_1} \tag{7}$$

Under assumptions regarding the regularity of the design matrix and a usual restricted eigenvalue condition (*i.e.* the minimal eigenvalue of the Gram matrix associated with the design matrix when the set of covariates is restricted to the non-zero coefficients is bounded away from zero), the author states results regarding the consistency properties of the Panel-Lasso. We can notice that this estimator is nothing more than a regular Lasso estimator with different penalty loadings for the coefficients of interest β and the unobserved heterogeneity c.

The assumption about the sparsity of c appears to be slightly odd specified in this way. Indeed, the economic thinking that underpins this assumption is that a large number of individuals belong to the main category for which $c_i = 0$ and only a few outliers, for which we cannot tell why they are outliers, exist. This assumption does not necessarily seem unrealistic, but seems at least not general enough. Firstly, if the unobserved heterogeneity is so anecdotal, an astute transformation of the model where the unobserved heterogeneity vanishes could do the job as well because we are seldom interested in estimating the unobserved heterogeneity if it is only some noise. Secondly, a case where we have grouped unobserved heterogeneity is a setting which would be more suitable for analysis and interpretability, and would also have the advantage of encompassing the Kock (2013a) model. Such a setting has been investigated by Bonhomme and Manresa (2012) in a small-dimensional case. Their "grouped fixed effects" estimator has the appealing property that individual group membership is data-driven and allows for nice interpretation of the division into groups.

In a nutshell, Kock (2013a)'s estimator is not an appealing solution because either we treat individual-specific heterogeneity as noise and find a way to do away with it, or we have an explanation regarding the source of this unobserved heteroegeneity in which case we want to be able interpret it in economic terms.

3.2 Belloni *et al.* (2014) solution: the Cluster Lasso

3.2.1 Presentation and intuition

In a very recent contribution, Belloni *et al.* (2014) propose a more convicing solution. They propose an estimator of β in model 3 that does not require the approximate sparsity assumption in the individual-specific heterogeneity c and that accounts for the fact that the data may be dependant within individual, heteroscedastic and non-Gaussian. They consider the "Within" version of model 3 where the c_i is eliminated by subtracting the mean of the variables:

$$\dot{y}_{i,t} = \dot{x}_{i,t}^T\beta + \dot{\epsilon}_{i,t} \tag{8}$$

Hastie *et al.* (2009, p. 61). The penalty we have just proposed corresponds indeed to a Normal prior distribution. The ℓ_1 penalty corresponds to a Laplace prior distribution.

with $\dot{z}_{i,t} = z_{i,t} - T^{-1}\sum_{t=1}^{T} z_{i,t}$. Then the so-called Cluster-Lasso estimator of the model is defined as:

$$\hat{\beta} \in \arg\min_{b} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\dot{y}_{i,t} - \dot{x}_{i,t}b)^2 + \frac{\lambda}{nT}\sum_{j=1}^{p}\hat{\phi}_j|b_j| \tag{9}$$

This program differs from the usual Lasso program in the sense that there are $p+1$ penalty terms: the main penalty level λ and covariate specific penalty loadings $\hat{\phi}_j$. Let us dwell a bit more on the use of the penalty terms $\hat{\phi}_j$. These penalty terms are needed to make sure that a so-called *"regularization event"* occurs with a high probability. This event is such that:

$$\frac{\lambda\hat{\phi}_j}{nT} \geq 2C\left|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{i,t,j}\dot{\epsilon}_{i,t}\right| \quad \forall j = 1, ..., p \tag{10}$$

Where C is a constant. Intuitively, this event is such that the penalty terms are large enough to dominate the noise associated with each regressor. If the noise, *i.e.* the correlation between the regressor and the true residual, is large, the penalty term $\hat{\phi}_j$ will be large enough so that $\hat{\beta}_j = 0$ as a result of program 9. We will not dive into how the probability of such an event is computed as it requires to go through moderate deviation theorems and takes us too far from econometrics. The ideal (infeasible) choice of penalty level is the following:

$$\phi_j = \sqrt{\frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\dot{x}_{i,t,j}\dot{\epsilon}_{i,t}\right)^2} \tag{11}$$

We can notice that it is a measure associated with the noise level of the j-th regressor. Since this optimal choice of penalty level relies on the unobserved residuals, the Cluster-Lasso is not feasible. However, the authors propose an iterative algorithm that makes the estimator computable in the spirit of the two-stage least squares for instrumental variables.

It is also to be noted that Lasso-type methods are often used only to perform variable selection. The nice feature of the ℓ_1 penalty is that some coefficients are set exactly to zero so we don't have to use a threshold to distinguish "real zeros" and "numerically small" coefficients. Then, a second step, called Post-Lasso, is used where the subset of variables which yield non-zero coefficients are kept into the design matrix and the model is re-estimated using ordinary least squares. This second step is of interest because it allows to do away with the bias (or shrinkage) associated with the ℓ_1 penalized regression.

3.2.2 Properties

Now, let us state and comment the three conditions under which we can tell that the Cluster-Lasso has good properties. These conditions are usual in statistical models that assume approximate sparsity. We will try to provide intuition for them instead of fulling stating them in a formal way (they can be found in the original paper).

Condition 1. *Approximately Sparse Model (ASM)*
Originally, in their paper Belloni et al. (2014) consider a non-parametric model of the form $y_{i,t} = f(w_{i,t}) + c_i + \epsilon_{i,t}$ *while we directly considered a linear model. This ASM condition states that this non-parametric model can be approximated by a linear combination of dictionary transformations of the original regressors* $w_{i,t}$*. This this why it is of interest to consider the case*

where $p >> n$ since we may want to consider a large dictionary to have a small approximation error of the true regression function.

Condition 2. *Sparse Eigenvalues (SE)*
This condition is very common in the high-dimensional literature. It appears to be natural in a sparisty context and allows to find faster rates of consistency for the Lasso when assumed. Let us denote by s the cardinal of the sparsity pattern. This condition states that with a probability approaching one as n tends to infinity, the empirical s-sparse Gram matrix has eigenvalues that are between two strictly positive constants κ and κ' that do not depend on n. We can notice that this condition is related to the singularity of the full Gram matrix when $p >> n$. In other words, in means that when the design matrix is restricted to the sparsity pattern, we no longer have a problem when inverting X^TX. For the usual Lasso, it amounts to assuming that the eigenvalues of the s-sparse matrix are bounded away from zero.

Condition 3. *Regularity Conditions (R)*
These are five conditions on the design matrix, penalty levels, residuals and sparsity pattern.

Under these conditions, we are now ready to state the main theorem of the paper regarding the feasible Custer-Lasso:

Theorem 1. *Model Selection Properties of Cluster-Lasso and Post-Cluster-Lasso*
Under conditions ASM, SE and R, for an overall penalty level given by $\lambda = 2C\sqrt{nT}\Phi^{-1}(1-\gamma/2p)$ and feasible penalty loadings such that $\ell\phi_j \leq \hat{\phi}_j \leq u\phi_j$ for some $\ell \to 1$ and $u \leq C < \infty$. Then the data dependent model $\hat{I}$ selected by a feasible Cluster-Lasso estimator satisfies with probability $1 - o(1)$, $\hat{s} = |\hat{I}| \leq Ks$ for some constant $K > 0$ that does not depend on n. In addition, we have the following relations for both the Cluster-Lasso and the Post-Cluster-Lasso:

1. $\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\dot{x}_{i,t}^T\hat{\beta} - \dot{x}_{i,t}^T\beta)^2 = \mathcal{O}_P\left(\frac{s\log(p\vee nT)}{n\iota_T}\right)$
2. $||\hat{\beta} - \beta||_2 = \mathcal{O}_P\left(\sqrt{\frac{s\log(p\vee nT)}{n\iota_T}}\right)$
3. $||\hat{\beta} - \beta||_1 = \mathcal{O}_P\left(\sqrt{\frac{s^2\log(p\vee nT)}{n\iota_T}}\right)$

It is to be noted that this result holds when either both n and T tend to $+\infty$ or only n tends to $+\infty$ with T fixed.

Theorem 1 shows that the feasible Cluster-Lasso has good variable selection properties, good predictions properties (relation 1) and good inference properties (relations 2 and 3). Relations 2 and 3 shows that the Cluster-Lasso and the Post-Cluster-Lasso have a high probability (on the regularization event) to converge in probability as n tends to infinity. In Appendix A we tried to provide a simple proof of the rate of convergence of the Cluster-Lasso. The proof of Belloni *et al.* (2014) is written in a complex way so we tried to make it simpler using the same steps as in the proof of Bickel *et al.* (2009) but did not manage to finish it.

4 Application

In this section we run estimations on simulated and real panel datasets in order to test and compare procedures in the high-dimensional case. All our estimations are first based on the Lasso, and corrected in a second step using post-Lasso estimation.

4.1 Simulated data

We test the Cluster-Lasso estimator proposed in Belloni *et al.* (2014) using simulated panel data. For this purpose we consider the model structure of 3 where we assume fixed effects and strong exogeneity, i.e. $\mathbb{E}[\epsilon_{i,t}|c_i, x_{i,1}, \dots, x_{i,T}] = 0$. The simulations are generated using the following DGP:

- ♢ High-Dimensional model: we set $(n, T, p) = (60, 5, 400)$ such that $n \times T < p$.
- ♢ Sparse structure of the true parameter: only a small number s of components are non null (and normalized to 1) in the true parameter β. In the following simulations we set $s = 10$.
- ♢ We generate the data in a very simple way:

$$x_{i,t}^j \sim \mathcal{N}(0, 4) \qquad \forall\,(i, t, j)$$

- ♢ We generate the residuals using an AR(1) specification:

$$\epsilon_{i,t} = \rho\epsilon_{i,t-1} + u_{i,t}, \quad u_{i,t} \sim \mathcal{N}(0, \sigma^2) \qquad \forall\,(i, t) \tag{12}$$

  where we set $\rho = 0.8$ and $\sigma = 0.5$.

- ♢ Correlation between fixed effects and covariates: the individual effects c_i are generated from one of the relevant covariates with the following (arbitrary) rule:

$$c_i = \frac{1}{T}\sum_{t=1}^{T} x_{1,t}$$

As a result we get values for the theoretically observed variables (y, X) and we can test the performance of the estimator previously presented. The full Data Generating Process can be summed up by the following equation

$$y_{i,t} = x_{i,t}^{'}\beta + c_i + \epsilon_{i,t} \tag{13}$$

Estimation is conducted with R using the package *penalized*. The *penalized()* function is flexible and allows us to solve directly the problem of interest:

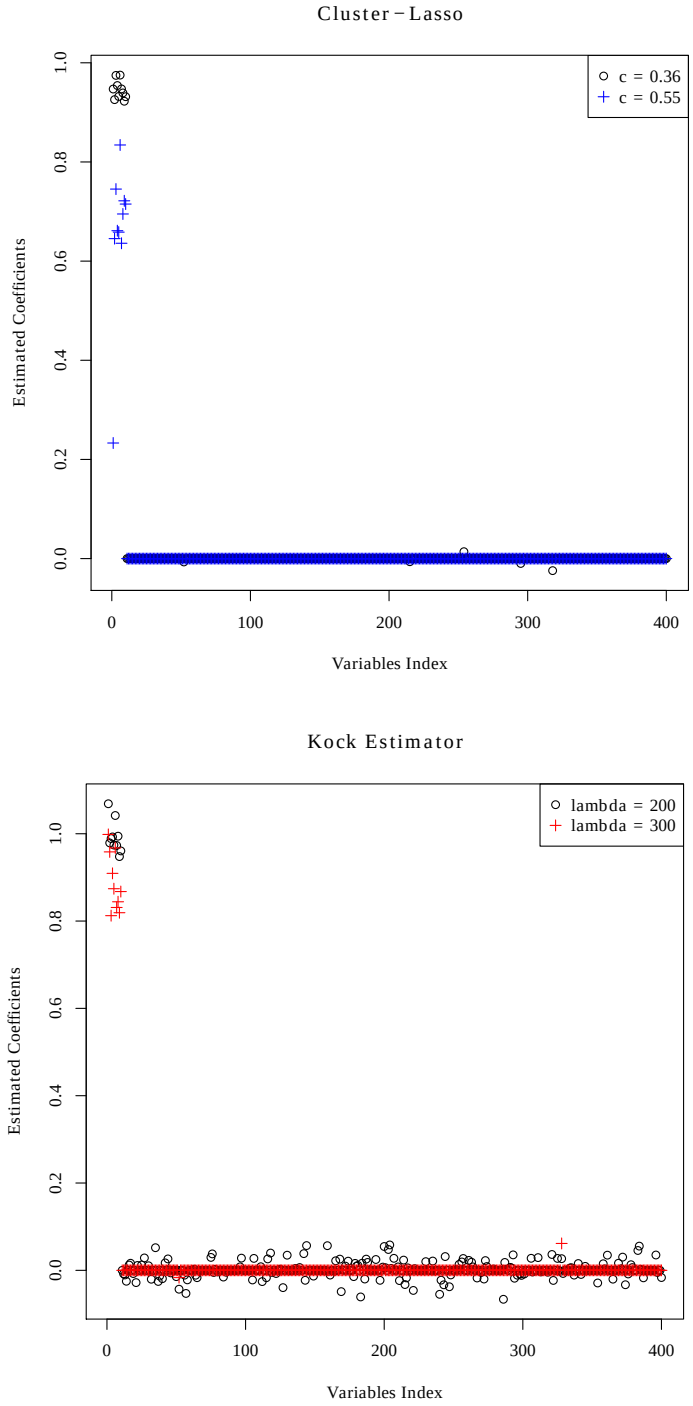
$$\hat{\beta} \in \arg\min_b \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\dot{y}_{i,t} - \dot{x}_{i,t}b)^2 + \frac{\lambda}{nT}\sum_{j=1}^{p}\hat{\phi}_j|b_j|$$

Note that in practice the minimization algorithm minimizes the sum of squares instead of the mean squares, so we can ignore the factor $\frac{1}{nT}$ in the penalty loadings. For our implementation we keep the Cluster-Lasso parameter values presented in the original reference for γ and λ, i.e. we set $\gamma = 0.1/\log(\min(p, nT))$ and $\lambda = c\sqrt{nT}\phi^{-1}(1 - \gamma/(2p))$. However in order to get a satisfying convergence we need to choose carefully and manually the value of the constant c. We

run estimation using our function implementing the algorithm (see the code in Appendix B). In a quite arbitrary maneer we set $c = 0.55$; with this value the algorithm converges and selects the correct variables in a few iterations ($K < 10$). Estimation results and comparison with Kock's estimator, that we also implemented, are presented in figure 1.

One could suspect that these good results come from a too simple Data Generating Process. In particular since our true parameter β only contains values in $\{0, 1\}$, the distinction between zero and nonzero coefficients might be easy to detect. In order to asses the robustness of the algorithm, we test the estimation procedure on real data in the following subsection.

Figure 1: Cluster - Lasso estimation on simulated dataset



4.2 Empirical example: another look at the union wage premium from Vella and Verbeek (1998)

We now turn to an application of the Cluster-Lasso to real data. We take a sample of wage data from Vella and Verbeek (1998) where the outcome variable is the log of the hourly wage. Data is taken from the National Longitudinal Survey (Youth Sample) and comprises a sample of full-time working males who have completed their schooling by 1980 and then followed over the period 1980 to 1987. There are 545 individuals because the authors have excluded those who drop out so as to obtain a balanced design. Description of the covariates can be found in Table 1 of Vella and Verbeek (1998). The aim of this article is to evaluate the union-wage premium, *i.e.* the incremental hourly wage that one gains from participating in a union. Unobserved heterogeneity induces the union participation variable to be endogenous because unobserved individual factors that influence the choice of participating in a union may be correlated to those which also determine the wage. Interestingly, Vella and Verbeek (1998) argue that unobserved heterogeneity is not only individual-specific, but also time-individual specific meaning that this unobserved heterogeneity changes over time. They find convincing evidence that introducing control functions that vary over time helps getting a better estimate of the union wage premium, as the coefficient associated with the union participation moves from .146 to .311 from the simple model estimated with OLS to the one with correction terms.

We take another look at this dataset and do not employ their methodology. Instead of considering a first-step estimation where the choice of being in a union or not is considered, we allow for a more general specification of the wage equation. Our assumption is that the bias in the estimation of the union wage premium may not come from individual-time specific unobserved heterogeneity but from a misspecification of the wage equation. For this purpose, we consider a high-dimensional design matrix and use the Cluster-Lasso to select the control variables. We do not penalize the coefficient associated to the union participation as it is the variable of interest.

There are only 33 covariates but we consider several transformations of them in order to consider a high-dimensional problem. This is of interest because the effect of some covariates may be non-linear and we want to be able to capture that effect. For each variable $x_{i,t}^j$ in the original data set, we introduce the new variables $(x_{i,t}^j x_{i,t}^{j'})$, $(x_{i,t}^j)^2$, and $(\log(x_{i,t}^j))$. We finally run our Cluster-Lasso estimation with $p = 305$ covariates.

As in the case of simulated data we need to manually adjust the value of the constant c to get nice selection properties. We will investigate several values for c and report the estimated coefficient associated with the union participation. The result is in Table 1.

Table 1: Size of the Union wage premium

Penalty (prop. to)	.32	.3	.275	.25	.2	.15	.1
Number of selected covariates	11	15	16	19	25	32	45
Union wage premium	.127	.141	.146	.080	.082	.074	.077

Note: this table reports the size of the union wage premium computed by Post-Cluster-Lasso and the number of variables selected by the Cluster-Lasso for different size of the penalty.

We also report the results of the Post-Cluster-Lasso for $c = .3$ in Table 2:

Table 2: Cluster-Lasso estimation of the wage equation, $c = 0.3$

	Cluster-Lasso Coefficient
EXPER	0.054102
HOURS	-0.000060
MAR	0.054139
UNION	0.140927
'EXPER x NE LOG'	0.104708
'EXPER x OCC2 LOG'	0.032371
'EXPER x S LOG'	0.102399
'EXPER x UNION LOG'	-0.036772
'HOURS x BLACK squared'	-0.000000
'HOURS x FIN'	0.000094
'HOURS x FIN squared'	-0.000000
'HOURS x S squared'	-0.000000
'HOURS x TRAD squared'	-0.000000
'SCHOOL x OCC4 squared'	-0.000396

As previously underlined the value of the constant c is critical in the estimation process. The Cluster-Lasso leads us to selecting the variables "experience", "married", and "union" as relevant to explain the log-wage. The variable "hours", and several transformations of the initial variables also seem to play an important role. Since Belloni *et al.* (2014) do not provide the asymptotic distribution of the Cluster-Lasso estimator in the general case, we can't run any hypothesis testing. However some variables such as years of schooling or the dummies "Black" and "Rural" are not selected while they are associated to significant, positive coefficients in the OLS estimation given by Vella and Verbeek (1998).

We find a coefficient of .141 associated to the covariate "union", a result close to the one in the reference article when they do not correct for the unobserved heterogeneity that also varies with time (.146). However, no matter the size of the penalty that we implement, we are not able to reproduce or even get close to a union wage premium of .3 that the authors found. This suggest that substantial individual-time specific unobserved heterogeneity may remain and that it biases the estimate of the union wage premium, as Vella and Verbeek (1998) suggested.

5 Conclusion and further roads

Methods to deal with high-dimensional problems are of interest in micro-econometrics mostly as a way to perform model selection, whether it is in a context of a non-parametric model estimated by sieve approximation, selection of control variables or instruments. We reviewed the particular case of fixed effect panel data models. These models do not conform to a straightforward application of the regular Lasso since the assumption of approximate sparsity in the individual-specific heterogeneity appears unrealistic and that temporal correlation must be taken into account especially when a Within transformation of the model is considered. A convincing estimator called Cluster-Lasso has been proposed by Belloni *et al.* (2014): it has desirable theoretical properties based on assumptions that are usual in the high-dimensional literature and is also computationally efficient.

However, this estimator suffers from several drawbacks. The first one is that the rule for the choice of the overall penalty level λ is vague. Some solutions have been proposed for example in Bickel *et al.* (2009) but we can notice this this quantity is crucial as it determines the amount of shrinkage chosen for the estimator. In empirical practice, several values for λ are often explored or it may also be chosen by cross-validation. Another concern is the practical implementation of the estimation procedure and the convergence of the algorithm. In particular Belloni *et al.* (2014) don't provide any precise rule concerning the choice of the number of steps K. In our experimentation a relatively small number of iterations has been enough to select a stable number of covariates.

Related to the idea that we have exposed regarding the relationship between the choice of the type of penalty and prior beliefs about the distribution of unknown coefficients, we can mention the penalized regressions and variable selection have also been investigated in a Bayesian framework, see for example Ročková and George (2014).

6 Bibliography

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999): "High Wage Workers and High Wage Firms". *Econometrica*, 67(2):251–333.

Angrist, J. D. and Krueger, A. B. (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics, MIT Press*, 106(4):979–1014.

Arellano, M. (2003): *Panel Data Econometrics*. Number 9780199245291 in OUP Catalogue, Oxford University Press. Oxford University Press.

Belloni, A. and Chernozhukov, V. (2009): "High Dimensional Sparse Econometric Models: An Introduction". In P. Alquier, E. Gautier, and G. Stoltz, editors, *Inverse Problems and High-Dimensional Estimation*. Springer Publishing Company, Incorporated.

Belloni, A. and Chernozhukov, V. (2010): "Post-l1-penalized estimators in high-dimensional linear regression models". CeMMAP working papers CWP13/10, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013): "Inference on treatment effects after selection amongst high-dimensional controls". CeMMAP working papers, Centre for Microdata Methods and Practice, Institute for Fiscal Studies CWP26/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2014): "Inference in High Dimensional Panel Models with an Application to Gun Control". *ArXiv e-prints*.

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

Bonhomme, S. and Manresa, E. (2012): "Grouped Patterns Of Heterogeneity In Panel Data". Working Papers, CEMFI wp20121208, CEMFI.

Cook, P. J. and Ludwig, J. (2004): "The Social Costs of Gun Ownership". NBER Working Papers 10736, National Bureau of Economic Research, Inc.

Hastie, T., Tibshirani, R., and Friedman, J. (2009): *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Springer, 2 edition.

Kock, A. B. (2013a): "Oracle Efficient Variable Selection In Random And Fixed Effects Panel Data Models". *Econometric Theory*, 29(01):115–152.

——— (2013b): "Oracle Inequalities for High-Dimensional Panel Data Models". *ArXiv e-prints*.

Koenker, R. (2004): "Quantile regression for longitudinal data". *Journal of Multivariate Analysis*, 91(1):74 – 89. Special Issue on Semiparametric and Nonparametric Mixed Models.

Lounici, K. (2008): "Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators". *Electronic Journal of Statistics*, 2:90–102.

Magnac, T. (2001): *Econométrie linéaire des panels: une introduction*. Document de travail - INSEE. INSEE.

Sala-i Martin, X. (1997): "I Just Ran Two Million Regressions". *American Economic Review, American Economic Association*, 87(2):178–83.

Meinshausen, N. and Yu, B. (2009): "Lasso-type recovery of sparse representations for high-dimensional data". *The Annals of Statistics*, 37(1):246–270.

Ročková, V. and George, E. I. (2014): "EMVS: The EM Approach to Bayesian Variable Selection". *Journal of the American Statistical Association*, 109(506):828–846.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Vella, F. and Verbeek, M. (1998): "Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men". *Journal of Applied Econometrics*, 13(2):163–183.

Wooldridge, J. M. (2001): *Econometric Analysis of Cross Section and Panel Data*. Number 0262232197 in MIT Press Books. The MIT Press.

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

Zhao, P. and Yu, B. (2006): "On Model Selection Consistency of Lasso". *J. Mach. Learn. Res.*, 7:2541–2563.

A Proof of the Cluster-Lasso

In this section, we tried to provide a proof of the rate of convergence of the Cluster-Lasso. The proof of Belloni *et al.* (2014) is written in a complex way so we tried to make it simpler using the same steps as in the proof of Bickel *et al.* (2009) but did not manage to finish it. We denote the objective function by $\mathcal{L}$, defined in the following way:

$$\mathcal{L}(\beta) := \hat{Q}(\beta) + \frac{\lambda}{nT}\sum_{k=1}^{p}\phi_k|\beta_k| \tag{14}$$

with $\hat{Q}(\beta) := (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{i,t} - x_{i,t}^T\beta\right)^2$. For a matter of simplicity, we dropped the "dot" superscript but we are dealing with the Within transforms of the variables. We call β^* the true value of the parameter. We also introduce the sparsity pattern of β^*: $J(\beta^*) := \left\{j : \beta_j^* \neq 0\right\}$. $\text{Card}(J) =\leq s$. By definition of the Lasso program, since we have a convex function, zero belongs to the sub-differential.

$$0 \in \partial\mathcal{L}(\hat{\beta}) \tag{15}$$

It means that there exist a p-dimensional vector u such that:

$$\frac{2}{nT}X^T(y - X\hat{\beta}) = \frac{\lambda}{nT}\phi u \tag{16}$$

$$\frac{2}{nT}\phi^{-1}X^T(y - X\hat{\beta}) = \frac{\lambda}{nT}u \tag{17}$$

Where we obtain the second inequality by multiply by the inverse of $\phi := \text{diag}(\phi_1, ..., \phi_p)$. Multiplying by a given β^T, and since $\forall j$, $|u_j| \leq 1$, we have:

$$\frac{2}{nT}\beta^T\phi^{-1}X^T(y - X\hat{\beta}) \leq \frac{\lambda}{nT}|\beta| \tag{18}$$

The same inequality holds true for $\hat{\beta}$:

$$\frac{2}{nT}\hat{\beta}^T\phi^{-1}X^T(y - X\hat{\beta}) \leq \frac{\lambda}{nT}|\hat{\beta}| \tag{19}$$

By taking one minus the other we obtain:

$$\frac{2}{nT}(\beta - \hat{\beta})^T\phi^{-1}X^T(y - X\hat{\beta}) \leq \frac{\lambda}{nT}(|\beta| - |\hat{\beta}|) \tag{20}$$

Replacing by the true model $y = X\beta + \epsilon$:

$$\frac{2}{nT}(\beta - \hat{\beta})^T\phi^{-1}X^TX(\beta^* - \hat{\beta}) \leq \frac{2}{nT}(\beta - \hat{\beta})^T\phi^{-1}X^T\epsilon + \frac{\lambda}{nT}(|\beta| - |\hat{\beta}|) \tag{21}$$

On the regularization event given by (2.3) in Belloni *et al.* (2014) we have:

$$\frac{1}{nT}|\phi^{-1}X^T\epsilon|_\infty \leq \frac{\lambda}{2cnT} \tag{22}$$

With $c > 1$. So we can notice that the first part of the right-hand side in equation 21 respects:

$$\frac{2}{nT}(\beta-\hat{\beta})^T\phi^{-1}X^T\epsilon \leq \frac{2}{nT}|\beta-\hat{\beta}|_1|\phi^{-1}X^T\epsilon|_\infty \tag{23}$$

$$\leq |\beta-\hat{\beta}|_1\frac{\lambda}{cnT} \tag{24}$$

And we have:

$$\frac{2}{nT}(\beta-\hat{\beta})^T\phi^{-\frac{1}{2}}X^TX\phi^{-\frac{1}{2}}(\beta^*-\hat{\beta}) \leq \frac{\lambda}{cnT}(|\beta-\hat{\beta}|+c|\beta|-c|\hat{\beta}|) \tag{25}$$

By developing the left-hand side we obtain:

$$\frac{1}{nT}||X\phi^{-\frac{1}{2}}(\hat{\beta}-\beta^*)||_2^2 \leq \frac{1}{nT}||X\phi^{-\frac{1}{2}}(\beta-\beta^*)||_2^2+\frac{\lambda}{cnT}(|\beta-\hat{\beta}|+c|\beta|-c|\hat{\beta}|)-\frac{1}{nT}||X\phi^{-\frac{1}{2}}(\beta-\hat{\beta})||_2^2 \tag{26}$$

We denote $\Delta=\hat{\beta}-\beta$ and we know that:

$$(|\beta-\hat{\beta}|+c|\beta|-c|\hat{\beta}|) \leq (1+c)|\Delta_J|+(1-c)|\Delta_{J^C}|+2c|\beta_{J^C}| \tag{27}$$

Now, we need to use the sparse eigenvalue (SE) condition to be able to consider several cases and provide the rate of convergence. We also need regularity conditions (R) to say something about the size of the norm of the penalty loadings ϕ. However, we are not able to continue the proof further.

B R code

```r
###################################
#### DATA GENERATING PROCESS #####
###################################

dgp_panel_data <- function(n, t, p, j, intercept = FALSE,
                           fixed_effects = TRUE, correlated_residuals =
                              FALSE)
  {
  set.seed(12071990)
  ## Unobserved residuals - Strong exogeneity asumption
  e <- rnorm(n = n*t, mean = 0, sd = 1) # uncorrelated component

  if (correlated_residuals == TRUE){
    rho <- 0.8
    e <- NULL
    for(i in 1:n){
      e <- c(e, AR(t,rho=rho))
    }
  }

  ## Explanatory variables
  indiv <- kronecker(rep(1:n), rep(1,t))
  X <- matrix(data = rnorm(n = (n*t)*p, mean=0, sd=2), ncol = p)

  ## Fixed effects

  c <- rep(0, n*t)
  if (fixed_effects == TRUE){
    Z <- data.frame(cbind(indiv, X))
    # hyp: the fixed effect is the mean of x_1 for each i
    #c <- tapply(Z$V3, Z$indiv, mean) + tapply(Z$V2, Z$indiv, mean)
    c <- tapply(Z$V2, Z$indiv, mean)
    c <- matrix(kronecker(c, rep(1,t)), ncol = 1)
  }

  ## True parameter with sparse structure
  beta <- c(rep(1, j), rep(0, p-j))
  var_labs <- c("indiv", "y", paste("V", rep(1:p), sep = ""))

  if (intercept == TRUE){
    X <- cbind(rep(1, n*t), X)
    beta <- c(1, beta)
    var_labs <- c("indiv", "y", '(Intercept)', paste("V", rep(1:p), sep =
        ""))
  }

  y <- X%*%beta + c + e

  sdata <- data.frame(cbind(indiv, y , X))
```

```
  colnames(sdata) = var_labs
  print(head(sdata))
  return(list(sdata = sdata, beta = beta, c = c, e = e))
}

AR <- function(T, rho=.8){
  x <- vector(length=T)
  x[1] <- rnorm(1,0,sd=2)
  for(t in 2:T){
    x[t] <- rho * x[t-1] + rnorm(1,0,sd=0.5)
  }
  return(x)
}
```

```
cluster_lasso <- function(data, lambda, K, post.lasso = TRUE){

  # Data: dataset with panel data structure, form is output of dgp_panel_data
  # lambda: overall penalty level
  # K: number of iterations

  library(plm)
  library(fBasics)
  library(lars)
  library(penalized)

  ## Compute the Within tranforms
  W <- within_matrix(data, index = 'indiv')
  y_dot <- W[,1]; X_dot <- W[,2:ncol(W)]

  ## Initialize the penalty loadings
  phi <- penalty_loadings(y_dot, X_dot, index=data$indiv)

  ## Start the loop
  for (i in 1:K){
    lasso <- penalized(response = y_dot, penalized = X_dot,
                       unpenalized = ~0, lambda1 = lambda  * diag(phi),
                          lambda2 =
                         0, model = "linear")
    # Do not divide penalty by n*T, since sum of square is minimized and not mean of squares
    selected_vars <- colnames(X_dot)[coef(lasso, which = 'all') != 0]
    if (post.lasso == TRUE){
      post_lasso <- lm(y_dot ~ ., data = as.data.frame(X_dot[,selected_vars]))
      e <- as.matrix(residuals(post_lasso)) # Recompute the residuals from a post-lasso
    }
    else{
      e <- as.matrix(residuals(lasso))
    }
```

```
    phi ← penalty_loadings(e, X_dot, index=data$indiv)
  }

  if (post.lasso == TRUE){
    return(list(coef = coefficients(post_lasso), residuals = e,
                names = names(coefficients(post_lasso)),
                coef_lasso = coef(lasso, which='all')))
  }
  else{
    return(list(coef = coef(lasso, which='all'), residuals = e,
                names = names(coef(lasso, which='all'))))
  }
}


within_matrix ← function(X, index = 'indiv'){
  J ← pdata.frame(as.data.frame(X), index = index,  drop.index = TRUE)
  W ← NULL
  for (i in 1:ncol(J)){
    W ← cbind(W, Within(J[,i]))
  }
  colnames(W) ← colnames(J)
  return(W)
}

g ← function(v,e,index){
  # Compute the penalty loading for a given regressor
  # Formula given in the article

  return(sqrt(sum((tapply(v*e,index,sum))^2)/length(e)))
}

penalty_loadings ← function(e, X, index){
  phi ← sapply(as.data.frame(X), g, e = e, index=index)
  return(diag(phi))
}
```