1 Introduction

The paper[1] presents a novel approach to identifying and estimating latent group structures in panel data, a crucial task for improving the accuracy and interpretiveness of models in economics, social sciences, and beyond. Latent structures[2] refer to the unseen or hidden relationships and patterns within complex data sets that are not directly observable. Recognizing them is vital because it enables the creation of more accurate and insightful models by revealing the underlying mechanisms that influence observed variables. This understanding can lead to more effective data interpretation and more precise predictions, and improve decision-making. Earlier studies typically handled heterogeneity in panel data by assuming complete slope homogeneity for specified parameters, modeling unobserved heterogeneity through individual-specific effects. This approach, while convenient, has often been questioned in empirical studies due to its inability to handle the heterogeneity in real-world data[3].

In this study, penalized techniques are introduced: Penalized Profile Likelihood (PPL) for linear and nonlinear models without endogenous regressors and Penalized GMM (PGMM) for linear models with endogeneity. The novel aspect of the methodology is the introduction of Classifier-Lasso (C-Lasso), a variant of the Lasso, which effectively shrinks individual coefficients to their group-specific values while simultaneously performing classification and estimation. This method stands out for its ability to consistently estimate and classify group membership in the presence of unknown group structures, achieving uniform consistency and, in some cases, an oracle property where the estimators are as effective as if the true group identities were known[4].

2 Methodological Overview

In the panel structure models, parameters are defined as follows:

1. $\alpha = (\alpha_1, \ldots, \alpha_{K_0})$: These are group-specific parameters. The number of groups K_0 is usually unknown in practice. Specifically, α_k represents the parameter vector for the k^{th} group. The group-specific parameters are what the individual-specific parameters β_i shrink towards in the penalized estimation process. The α parameters define the commonalities within groups and differences across them.
2. $\beta = (\beta_1, \ldots, \beta_N)$: These are individual-specific parameters. Specifically, β_i represents the parameter vector for the i^{th} individual in the panel data. These parameters are assumed to follow a group pattern, indicated as:

$$\beta_i^0 = \alpha_k^0 \quad \text{for} \quad i \in G_k^0$$

[1] Liangjun Su, Zhentao Shi, and Peter C. B. Phillips. "Identifying Latent Structures in Panel Data". In: *Econometrica* 84.6 (2016), pp. 2215–2264.

[2] *Latent and Observable Variables*. In: *Wikipedia*. Oct. 3, 2023.

[3] Su, Shi, and Phillips, see n. 1, p. 2216.

[4] The oracle property is a desirable feature of a variable selection method in high-dimensional statistical modeling. It means that the method can consistently select the correct subset of predictors and estimate the coefficients of the non-zero predictors as well as if the true model were known in advance. In other words, the method can perform as well as an oracle who knows the true model.See Jianqing Fan and Runze Li. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties". In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1348–1360. JSTOR: 3085904; Richard Hardy. *What Is the Oracle Property of an Estimator?* Cross Validated. Jan. 27, 2017. URL: https://stats.stackexchange.com/q/229142 (visited on 12/26/2023)

where G_k^0 is the specific group to which individual i belongs, and k indexes the group[5].

In this study, $\psi(w_{it};\beta_i,\mu_i)$ represents the negative logarithm of the pseudo-true conditional density function.

- This function is defined for the variable y_{it} conditional on x_{it}, the history of both (y_{it},x_{it}), and the parameters (μ_i,β_i).
- μ_i denotes scalar individual effects, which are specific to each individual in the study.
- β_i refers to a $p\times 1$ vector of individual-specific parameters of interest.

2.1 Penalized Profile Likelihood (PPL) Method

Following Hahn and Newey (2004) and Hahn and Kuersteiner (2011), the profile log-likelihood function is defined for a given individual i and time t as follows:

$$Q_{1,NT}(\beta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\psi(w_{it};\beta_i,\hat{\mu}_i(\beta_i))$$

where $\hat{\mu}_i(\beta_i)=\arg\min_{\mu_i}\frac{1}{T}\sum_{t=1}^{T}\psi(w_{it};\beta_i,\mu_i)$. Here, $\psi(w_{it};\beta_i,\mu_i)$ represents a specified function of the data w_{it}, individual parameters β_i, and μ_i. It can differ in different cases, according to the model assumption.

The goal is to estimate β and α by minimizing the following Penalized Profile Likelihood (PPL) criterion function:

$$Q_{1\,NT\lambda_1}^{(K_0)}(\beta,\alpha)=Q_{1,NT}(\beta)+\frac{\lambda_1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K_0}\|\beta_i-\alpha_k\|$$

where $\lambda_1=\lambda_{1NT}$ is a tuning parameter. The minimization of this criterion function produces classifier-Lasso (C-Lasso) estimates $\hat{\beta}$ and $\hat{\alpha}$ of β and α, respectively.

Traditional Lasso methods incorporate an additive penalty term that shrinks coefficients toward zero to enforce sparsity and variable selection. The C-Lasso, however, includes both additive and multiplicative penalty components. This mixed form allows each individual parameter β_i to be shrunk towards one of several possible group-specific parameters α_k.

Unlike traditional Lasso, which is generally used for individual parameter shrinkage and selection, the C-Lasso is designed to identify and differentiate between multiple groups within the data. It does so by allowing each parameter to shrink toward multiple potential group-level parameters, effectively classifying the data points into different groups based on their characteristics.

Under the assumptions in Section 2.3, the PPL C-Lasso estimators have the following asymptotic properties.

Theorem 2.1 (Consistency and Convergence Rates):
Assumptions Required: Assumption A1, along with the condition $\lambda_1=o(1)$.
Property: Pointwise and mean square convergence of individual parameter estimates $\hat{\beta}_i$ and the consistency of group-specific parameters $\hat{\alpha}_k$.

Theorem 2.2 (Uniform Consistency for Classification):
Assumptions Required: Assumptions A1 and A2.

[5] The superscript 0 is used to denote the true values

Property: Uniform consistency in classification, ensuring individuals are classified into the correct groups with high probability as sample size increases.

Corollary 2.3 (Consistency of Group Sizes):
Assumptions Required: Assumptions A1 and A2.
Property: Consistency of the estimated number of individuals in each group, $\hat{N}_k$, for the true number N_k.

Theorem 2.4 (Oracle Property):
Assumptions Required: Assumptions A1, A2, and A3.
Property: Oracle property of the PPL estimator $\hat{\alpha}_k$ and the asymptotic normality of the distribution of $\sqrt{N_k T}(\hat{\alpha}_k - \alpha_k^0)$.

Theorem 2.5 (Asymptotic Distribution of Post-Lasso Estimators):
Assumptions Required: Assumptions A1, A2, and A3.
Property: Asymptotic distribution of Post-Lasso estimators[6] , indicating that they are asymptotically equivalent to the PPL estimators with similar distributional properties.

2.2 Penalized GMM Estimation

In Section 3, the authors concentrate on estimating the parameters α and β within the context of a linear panel structure model represented by equation (2.3). Specifically, they consider the first-differenced system:

$$\Delta y_{it} = \beta_i^{0\prime} \Delta x_{it} + \Delta \epsilon_{it}.$$

The goal is to minimize the PGMM criterion function:

$$Q_{2NT,\lambda_2}^{(K_0)}(\beta, \alpha) = Q_{2,NT}(\beta) + \frac{\lambda_2}{N} \sum_{i=1}^{N} \prod_{k=1}^{K_0} \|\beta_i - \alpha_k\|$$

Here:

- $\lambda_2 = \lambda_{2NT}$ is a tuning parameter.
- $Q_{2,NT}(\beta) = \frac{1}{N}\sum_{i=1}^{N} \left[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it})\right]^T W_{i,NT} \left[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it})\right]$
- $W_{i,NT}$ is a symmetric matrix that becomes asymptotically nonsingular.

Minimizing this criterion function yields the PGMM estimates $\tilde{\alpha}$ and $\tilde{\beta}$, where $\tilde{\alpha}$ is defined as $\tilde{\alpha} \equiv (\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_{K_0})$ and $\tilde{\beta}$ is defined as $\tilde{\beta} \equiv (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_N)$.

The PGMM estimators exhibit specific asymptotic properties under certain assumptions. These properties also include consistency, where estimators converge to the true parameters, and asymptotic normality, indicating that the distribution of estimators approaches a normal distribution as the sample size grows. Classification consistency ensures that the probability of misclassifying individuals into incorrect groups diminishes. The post-Lasso estimators, a refinement of the initial estimates, also demonstrate improved convergence rates and asymptotic normality.

[6] Post-Lasso estimators are a two-stage procedure used in regression to handle high-dimensional data. Initially, Lasso regression is employed to select a subset of relevant variables by shrinking some coefficients to zero. This reduces model complexity and mitigates overfitting. In the subsequent step, ordinary least squares (OLS) regression is applied but only on the variables selected by Lasso. This approach aims to combine Lasso's variable selection capability with the unbiased estimation of OLS, offering a balance between model accuracy and interpretability, especially beneficial in settings with many predictors but relatively few significant ones. See Alexandre Belloni and Victor Chernozhukov. "Least Squares after Model Selection in High-Dimensional Sparse Models". In: *Bernoulli* 19.2 (2013), pp. 521–547. JSTOR: 23525734

3 Simulations and Empirical Applications

In the simulations, the finite-sample performance of their proposed methods for panel structure models with unknown group membership is evaluated. The study considers three data generating processes (DGPs) that differ in the model specification, the number of groups, and the parameter values. It varies the sample size, the tuning parameters, and the number of candidate groups in the simulations. It finds that their methods achieve high accuracy, and are robust to different choices of tuning parameters[7].

Section 5.1 highlights C-Lasso's robustness in empirical analysis, applying it to savings rates and civil war incidence across countries. In the savings rate study, it assesses macroeconomic indicators like inflation and GDP growth for 56 countries, revealing the importance of homogeneity in panel data analysis. The results, indicating significant variations especially in the impact of inflation and interest rates, underscore the method's ability to discern distinct economic patterns. In the civil war study, it classifies 38 countries based on incidence rates, uncovering a negative relationship between GDP and civil war in less affected countries. These applications emphasize C-Lasso's value in economic research。

4 Subsequent Developments and Related Works

4.1 Lu and Su 2017

In Section 2.5 and 3.4, the 2016 study provides a method for determining the number of groups as part of its broader objective to identify and estimate latent structures. It introduces a selection method based on minimizing an information criterion (IC). It also outlines specific assumptions required for the consistency and accuracy of this method and provides a theorem supporting the effectiveness of the information criterion in selecting the correct number of groups under the specified conditions.

In the following 2017 study[8], the authors specifically focus on the empirical determination of the number of groups in **linear** latent panel structures. While the study is built on the framework established in the 2016 study, it delves deeper into the specification testing aspect of the problem. It proposes a specific testing procedure (a residual-based Lagrange multiplier-type test) for determining the number of groups. This approach involves setting up hypotheses about the number of groups and then testing these hypotheses empirically to ascertain the correct number. Researchers could also first use the LM-type test to determine the number of groups empirically and then apply the C-Lasso method for classification and parameter estimation within this determined group structure.

4.2 Su and Ju 2018

As remarked in the 2016 paper, this paper[9] extends the C-Lasso approach to dynamic panel data models with interactive fixed effects (IFE), emphasizing the estimation of latent grouped patterns. It addresses the complexity introduced by cross-section dependence and the dynamic nature of the models. The paper introduces penalized principal component (PPC) estimation to handle the challenges posed by interactive

[7] I don't understand the choice of tuning parameters.

[8] Xun Lu and Liangjun Su. "Determining the Number of Groups in Latent Panel Structures with an Application to Income and Democracy". In: *Quantitative Economics* 8.3 (2017), pp. 729–760.

[9] Liangjun Su and Gaosheng Ju. "Identifying Latent Grouped Patterns in Panel Data Models with Interactive Fixed Effects". In: *Journal of Econometrics*. Special Issue on Advances in Econometric Theory: Essays in Honor of Takeshi Amemiya 206.2 (Oct. 1, 2018), pp. 554–573.

fixed effects in the panel data models. This methodology is a progression from the earlier C-Lasso, incorporating additional considerations for the dynamic aspects and cross-sectional dependencies of the data. It continues to emphasize uniform classification consistency and the oracle property of the estimators. It offers simulations to demonstrate the finite-sample performance and applies the improved method to study housing prices in China, identifying latent groups based on price persistence patterns.

4.3 Su et al. 2019

The 2019 study[10], progressed by proposing a heterogeneous time-varying panel data model with latent group structures that allows coefficients to vary over both individuals and time. This model treated coefficients as smooth functions of time and proposed a penalized sieve estimation method based on the classifier-Lasso (C-Lasso) for identifying individual membership and estimating group-specific functional coefficients. This improvement addressed the need for modeling individual heterogeneity and smooth structural changes over time, which are often observed in longitudinal or panel data sets. The 2019 study also extended the application range, demonstrating the approach's effectiveness in classifying and estimating in different contexts and providing more robust and flexible modeling for panel data with latent structures.

4.4 Wang and Su 2021

The 2021 study[11] proposes a procedure for identifying latent group structures in nonlinear panel data models. This includes a sequential binary segmentation algorithm (SBSA) for estimating group structures and a detailed examination of the model parameters' estimation and their asymptotic properties. It allows the presence of common parameters across all individuals, corresponding to the mixed panel structure model mentioned in Section 2.7 of the 2016 study.

5 Conclusion

The review systematically evaluates the Classifier-Lasso (C-Lasso) method's innovative approach in identifying and estimating latent group structures in panel data. It underscores the method's ability to handle heterogeneity and classify data into coherent groups, which is vital for enhancing the interpretability and precision of econometric models. The paper's exploration through simulations and empirical applications demonstrates the method's robustness and potential in various economic contexts . Related works and continuous developments show that the group of researchers are keeping focused on the topic and the C-Lasso could become more widely used, indicating its important influence on the future of econometric research and methods. Apart from the additional research proposed in the papers, the choice of tuning parameters may be vital for the C-Lasso method and remains an area ripe for further study.

[10] Liangjun Su, Xia Wang, and Sainan Jin. "Sieve Estimation of Time-Varying Panel Data Models With Latent Structures". In: *Journal of Business & Economic Statistics* 37.2 (Apr. 3, 2019), pp. 334–349.

[11] Wuyi Wang and Liangjun Su. "Identifying Latent Group Structures in Nonlinear Panels". In: *Journal of Econometrics*. Annals Issue: Celebrating 40 Years of Panel Data Analysis: Past, Present and Future 220.2 (Feb. 1, 2021), pp. 272–295.