Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis

Econometrics and Operations Research

Constructing a Lasso Based Approach for the Estimation of PSTR Models with Latent Group Structures

Author:	Meike Pluym (466476)
Supervisor:	W. Wang
Second assessor:	M. Grith

July 5, 2020

Abstract

This research proposes a novel mechanism for the incorporation of heterogeneity in panel data. The mechanism uses a Lasso based estimation of Panel Structure Threshold Regression (PSTR) models with latent group structures. This new method incorporates the heterogeneity of data into the model in two ways, introducing thresholds and allowing the coefficients to exhibit group-dependency. I evaluate the model by use of several Monte Carlo simulations which allow for the identification of various fit measures to show the performance of the model. Ultimately, the simulation shows promising results. As the time dimension of the panel data increases, the probability of identifying the true number of groups approaches 1, less than 1% of individuals is allocated to the wrong group, and the estimated coefficients approach the true values of the coefficients with shrinking biases. In addition, the applicability of the algorithm is shown by use of an empirical data set concerning a panel data set of savings rates of 55 countries.

Contents

1 Introduction

Can we estimate all observations in a data set using the same econometric model? Are parameters identical across time and cross-section dimensions of data? In recent years, the growth of widespread data collection has called for the identification of underlying group structures in data sets. With the growing emphasis on differences between individuals, firms, and countries, the importance of incorporating these differences into econometric models has too. While this importance has grown, literature on panel data models with unobserved group effects has increased greatly as well (e.g. Lin and Ng (2012), Ando and Bai (2015)). Along with the incorporation of unobserved group effects in data, research into assigning group membership on an observable threshold has become more popular (e.g. Chan (1993), Hansen (1999)). The introduction of unobserved group structures and observed threshold variables into a model can incorporate the heterogeneity in data into a model. By doing so, generally, estimation becomes more consistent and the fit of the model improves (Smith (1956)).

While both grouping algorithms and threshold models have been investigated separately to a great extent, there is very little documentation on the combination of the two. In terms of the latent group structures in data, K-means-like algorithms (see Bonhomme and Manresa (2015)), Finite Mixture Models methods (see McLachlan and Peel (2006)), and Lasso methods (see Su et al. (2016)), as well as many others have been studied extensively in past research. Threshold models, however, have fewer documentation. Although their relevance and appeal have been shown by multiple researchers (e.g. Hansen (1999)), econometric techniques for these type of models are investigated to only a small extent.

In previous research, Miao et al. (2020) propose an algorithm which combines a K-means-like clustering method with a threshold estimation algorithm in order to establish the estimation of a threshold regression model with latent group-structures. The researchers estimate group-membership by minimising the sum of squared errors of the model given the group-membership of each individual. The research shows very promising results and confirms the benefit we can gain from combining grouping techniques with threshold models.

As previously mentioned, as far as literature goes, there is no further investigation into incorporating the heterogeneity in a model using both grouping techniques and a threshold structure in the model. This research, therefore, aims to introduce a new method by constructing a linear threshold model which includes latent group structures in the data. I propose an algorithm which extends the C-Lasso clustering algorithm proposed by Su et al. (2016), to incorporate threshold effects in the data. This research is based on the ideas proposed by Miao et al. (2020). However, it takes a different approach for the estimation of the model. By conducting this research, I expect to show the performance of a Lasso based technique in a threshold environment and propose a new method for the incorporation of heterogeneity of a data set into the econometric model.

The performance of this new technique is established by performing a Monte Carlo simulation. The simulation allows for the evaluation of various fit measures of the model. In addition, a set of panel data on savings rates of 55 countries is evaluated using the estimation algorithm to show its applicability.

The paper is structured as follows. First, in Section 2, I discuss previous literature on the estimation of threshold regressions with and without taking into account the latent group structures in data, and some research on Lasso based estimation techniques. Then, Section 3 introduces and explains the estimation algorithm, discussing each step of the algorithm in detail. After, the simulation procedures are discussed in Section 4. This then leaves the discussion of the results of the simulation (Section 5) and the discussion of the empirical data set and its results (Section 6). Lastly, I conclude the paper by giving an overview of the results and proposing some areas for possible future research.

2 Theoretical framework

In order to investigate the performance of the Lasso clustering method in panel threshold regressions, this research focuses and extends upon the models proposed by Hansen (1999), Miao et al. (2020), and Su et al. (2016). Hansen (1999) estimates a panel threshold regression without taking into account possible latent group structures in the data. Miao et al. (2020) extend on these views by proposing a model which incorporates a K-means type clustering technique into a panel threshold regression model. Lastly, Su et al. (2016) propose the Classifier-Lasso (C-Lasso) technique which allows for the clustering of a set of panel data, without taking into account possible threshold effects. This C-Lasso technique achieves classification of group-specific estimators simultaneously in one single step.

The goal of this research is to extend the C-Lasso technique with elements from the panel threshold regression clustering algorithm in order to create a Lasso approach for constructing group-specific estimators for a panel threshold regression. In the next subsections, first the concept of

Panel Structure Threshold Regression (PSTR) models is discussed. Then, I focus on the research conducted by Miao et al. (2020). Lastly, I explain and discuss the concept of Lasso based estimation approaches.

2.1 Panel structure threshold regressions

This paper considers the clustering of a PSTR model. The threshold coefficient in such a model takes a value used to distinguish ranges of the threshold variable, where the behaviour above and below the threshold coefficient vary in an important manner. Huang et al. (2005), for example, show the presence of threshold effects in the modelling of oil price change and its volatility on economic activities. They show that if oil price change is below some threshold, the oil price change and volatility have limited impact on economies. While for values above the threshold, these variables explain the model well. Hansen (1999) shows the performance of thresholds in non-dynamic panel data with individual-fixed effects. Caner and Hansen (2004) extend upon Hansen's earlier work and investigate instrumental variable estimation of a threshold regression model with endogenous variables and an exogenous threshold. Thereafter, Kremer et al. (2012) extend on this research in order to allow estimation for dynamic panels, as well as static panel data. In addition, Girma (2005) shows the presence of a threshold effect in the correlation between foreign direct investment and productivity growth. All in all, threshold regression models are of importance in literature.

The general form of PSTR models is

$$y_{it} = \mu_i + \beta' x_{it} + \delta' x_{it} \mathbf{1}\{q_{it} > \gamma\} + \epsilon_{it}, \tag{1}$$

where q_{it} is the threshold variable and is, usually, predetermined. Furthermore, γ is the threshold coefficient.

Generally, the estimation of the threshold coefficient is based on the minimisation of some criterion function given the threshold, γ. Hansen (1999) proposes, for example, that the threshold coefficient can be estimated by minimising the sum of squared errors of the model given the value for γ. This method is a popular manner of estimating the threshold. In this research, I use the theories proposed by Hansen (1999) in order to estimate the threshold coefficients.

2.2 PSTR using K-means

In recent research, Miao et al. (2020) propose a K-means based algorithm which allows for the estimation of PSTR models with latent group structures. The researchers propose a method which takes into account these latent group structures in a data set by estimating group-specific slope, and threshold-effect coefficients as well as group-specific threshold coefficients. The method is based upon updating the group-membership, slope, and threshold-effect coefficients and threshold coefficients in each iteration until numerical convergence. This updating of coefficients and group-memberships is based on obtaining a minimal Sum of Squared Errors (SSE) of the model.

The researchers define an SSE function, $Q(\Theta, \mathbf{D}, \mathbf{G})$, where $\mathbf{D} \equiv (\gamma_1, ..., \gamma_G)' \in \Gamma^G$ is the set of group-specific threshold values, $\mathbf{G} \equiv (g_1, ..., g_N)' \in G_N$ is the set of group memberships of the individuals, $\Theta \equiv (\theta'_1, ..., \theta'_G)' \in B^G$ is the set of group-specific parameters, and $\theta_g \equiv (\beta'_g, \delta'_g)' \in B \subset \mathbb{R}^{2K}$ incorporates both the slope and the threshold parameter. The SSE minimisation is presented as follows

$$(\hat{\Theta}, \hat{\mathbf{D}}, \hat{\mathbf{G}}) = \underset{(\Theta,\mathbf{D},\mathbf{G}) \in B^G \times \Gamma^G \times G^N}{\arg\min} Q(\Theta, \mathbf{D}, \mathbf{G}) = \underset{(\Theta,\mathbf{D},\mathbf{G}) \in B^G \times \Gamma^G \times G^N}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} [\tilde{y}_{it} - \tilde{z}_{it}(\gamma_{g_i})' \theta_{g_i}]^2, \quad (2)$$

where $\tilde{z}_{it}(\gamma_{it}) = (\tilde{x}'_{it}, \tilde{x}'_{it}(\gamma))$, such that it incorporates the threshold variables present in the model. The optimisation of this model calls for the optimisation of the threshold value, group membership, and slope, and threshold parameters. To accomplish this, an algorithm is constructed such that first the $\hat{\mathbf{D}}$ is estimated using a known $\mathbf{G}$ and then, for a given $\mathbf{D}$ and $\mathbf{G}$, the slope and threshold parameters are estimated.

Miao et al. (2020) use a K-means like method to estimate $\mathbf{G}$. The step in their algorithm which allows for the updating of group-membership is

$$g_i = \underset{g \in \mathbf{G}}{\arg\min} \sum_{t=1}^{T} [\tilde{y}_{it} - \tilde{z}_{it}(\gamma_g)' \hat{\theta}_g]^2. \quad (3)$$

The benefits of using this method for group-membership estimation include the fact that it does not include any major computations, making the method simple and quick. In addition, this type of method does not require the identification of, for example, distributions in the data a priori. Merely the number of groups calls for a priori identification.

Nevertheless, the method has downsides. The method is sensitive to initialisation values. Depending on the initial group-allocations defined, the method may result in varying final clusters for different runs of the algorithm. Furthermore, the number of clusters has to be defined a priori for the algorithm to work. This causes a restriction on the freedom of the estimation method. This restriction can potentially be solved by incorporating the estimation of number of groups into the estimation algorithm. A method which allows this is a Lasso based estimation method.

2.3 The Lasso clustering technique

In this investigation, I propose the estimation of PSTR models using a Lasso type estimation procedure. The Least Absolute Selection and Shrinkage Operator (Lasso) technique is one of the best known and well documented estimation techniques. This method, first proposed by Tibshirani (1996), estimates models by minimising the error of a model subject to the sum of the absolute values of the coefficients being less then some constant. As shown by Tibshirani (1996), this method uses properties from both the ridge type regression and subset selection which allow for the estimation of coefficients in a linear model. Later, Tibshirani et al. (2005) proposed the fused Lasso method which, in extension to the Lasso method, takes into account potential structures in the data that have to be taken into account in the analysis of the model. While this method takes into account structures in data, it cannot identify latent group structures in the data. This is due to the fact that, generally, there is no natural structure between individuals present in a data set. In order to solve this, Su et al. (2016) propose an alternate version of the Lasso method which is able to identify these group structures explicitly, called the Classifier-Lasso (C-Lasso) technique.

In this C-Lasso model, it is assumed that the true values of the slope parameter follow group patterns as follows

$$\beta_i = \sum_{k=1}^{K} \alpha_k \mathbf{1}\{i \in G_k\} \qquad \text{for } i = 1, ...N,\ k = 1, ..., K, \tag{4}$$

where α take the group-specific coefficient values of group G_k. Su et al. (2016) introduce a multiplicative penalty term, instead of the additive as proposed in previous research (e.g. Tibshirani (1996)). Because of the multiplicative form of the penalty term, each slope parameter is allowed to shrink to a specific, a priori unknown, group parameter.

The penalty term in this model takes an additive form over all individuals, while taking a multiplicative form over all groups for each individual, such that the estimation of the model calls for the minimisation of the error in the model as well as the following error term

$$Error = \frac{\lambda_1}{N} \sum_{i=1}^{N} \prod_{k=1}^{K_0} \|\beta_i - \alpha_k\| \tag{5}$$

. This Lasso method is investigated extensively in literature, and researchers such as Tibshirani et al. (2005), Su et al. (2016), Chan et al. (2014), and Poetscher and Leeb (2007) have shown its relevance and performance in both the grouping of data and obtaining consistent estimates.

This method, combines the two steps concerning the estimation of slope, and threshold-effect coefficients, as proposed by Miao et al. (2020), into one step. In addition, this estimation algorithm of coefficients and group-memberships does not require a priori allocation of the value for the number of groups in the data set. Rather, it estimates the optimal group-structure in the data with corresponding coefficients simultaneously. The main benefit of this cluster estimation technique, as described by Su et al. (2016), is that it does not require any specification of modeling mechanisms for the group-structures. It is completely data-determined.

The disadvantage of using this technique to identify the latent group-structures and coefficients in the estimation of PSTR models is that it requires very expensive calculations. This causes a time-intensive estimation technique. Nevertheless, this paper uses the theories proposed by Su et al. (2016), namely the C-Lasso algorithm, to cluster the data and obtain estimates for the group-specific coefficients.

3 Methodology

The model under consideration in this research is a PSTR model. The model, in addition to the (heterogeneous) threshold, contains group-specific coefficients. In the following subsections, I present the model and explain the estimation technique. In Section 3.1, I present the PSTR model under consideration and give a global overview of the estimation algorithm. Then, in Section 3.2, I explain the step of the algorithm dedicated to the estimation of the threshold coefficient. Lastly, in Section 3.3, I explain the use of the Lasso technique which estimates the group-membership, slope and threshold-effect coefficients.

3.1 The panel threshold model

The general threshold model with latent group structures under consideration takes the following form

$$y_{it} = \mu_i + x_{it}'\beta^0_{g_i^0} + x_{it}'\delta^0_{g_i^0} \cdot \mathbf{1}\{q_{it} > \gamma^0_{g_i^0}\} + \epsilon_{it}, \qquad i = 1, ..., N,\ t = 1, ..., T, \tag{6}$$

where $i \in \{1, 2, ..., N\}$ represents the individuals in cross-section dimension, N, of the panel, and $t \in \{1, 2, ..., T\}$ the time periods in the time dimension, T. Furthermore, μ_i contains fixed individual effects of the model and ϵ_{it} the idiosyncratic error term. x_{it} is the $p \times 1$, where p is the number of regressors, vector of exogenous independent variables, and q_{it} the scalar threshold variable. As for the parameters, the model contains different "types". β_{g_i} and δ_{g_i} are both $p \times 1$ vectors of slope and threshold-effect coefficients, respectively. Each β_{g_i} and δ_{g_i} lie in space B. The scalar threshold coefficient, γ_{g_i} is defined in the space $[\underline{\gamma}, \overline{\gamma}]$, the definition of which is given in a later section. Furthermore, it is important to note that the slope, threshold-effect, and threshold are all group specific, i.e., they are dependent on g_i. This g is the group-membership variable and is defined in the space $g \in \{1, 2, ...G\}$, where, for the number of groups G, $\cup_{k=1}^{G^0} g_k^0 = \{1, 2, ..., N\}$ and $G_k^0 \cap G_j^0 = \emptyset$, for $j \neq k$. As for notation, superscript "0" in $\beta^0_{g_i^0}$, $\delta^0_{g_i^0}$, $\gamma^0_{g_i^0}$, and g_i^0, denote the true values of the coefficients.

In the estimation procedure, I use various notations for the collection of parameters. As each group has its own set of parameters, I define the total set of parameters to be $\mathbf{D} \equiv (\gamma_1, ..., \gamma_G)' \in \Gamma^G$, $\mathbf{G} \equiv (g_1, ..., g_N)' \in G^N$, $\Theta \equiv (\theta_1', ..., \theta_G')' \in B^G$, where $\theta_g \equiv (\beta_g', \delta_g')'$. Using this definition of θ, I define the totality of the independent variables as $x_{it}'(\gamma) = (x_{it}', x_{it}' \cdot \mathbf{1}\{q_{it} > \gamma\})$, such that we can use the short-hand notation $x_{it}'(\gamma)\theta$ to describe both the slope and threshold-effect and value.

In order to simplify the estimation, I choose to remove the individual fixed effect, μ_i, from the model. To do this, the mean over time for each individual, i, is subtracted from the value of both the dependent, y_{it}, and independent, x_{it}, variables, as well as the error term. The resulting dependent variable takes the form $\tilde{y}_{it} = y_{it} - T^{-1}\sum_t^T y_{it} = y_{it} - \overline{y}_i$. The independent variable and error term are defined analogously, such that $\tilde{x}_{it} = x_{it} - \overline{x}_i$, and $\tilde{\epsilon}_{it} = \epsilon_{it} - \overline{\epsilon}_i$. The resulting model is defined as

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta^0_{g_i^0} + \tilde{x}_{it}'\delta^0_{g_i^0} \cdot \mathbf{1}\{q_{it} > \gamma^0_{g_i^0}\} + \tilde{\epsilon}_{it}, \qquad i = 1, ..., N,\ t = 1, ..., T \tag{7}$$

The resulting model is estimated by an algorithm with two explicit steps; the estimation of the group-memberships and corresponding slope and threshold-effect coefficients, using the C-Lasso method proposed by Su et al. (2016), and the estimation of the scalar threshold value, based on the method proposed by Hansen (1999). The algorithm is constructed such that it alternates between re-estimating the threshold variable and accordingly updating the group-membership and coefficient estimates until numerical convergence. An overview of the algorithm can be seen below.

The estimation algorithm:
Input: the data, the maximum number of groups, the possible values for the tuning parameter.

1. Initialise the algorithm by assigning initial values for the coefficients
2. Estimate the threshold coefficient using theories proposed by Hansen (1999)
3. Estimate the group-membership, slope, and threshold-effect coefficients, using theories proposed by Su et al. (2016)
4. Repeat steps 2 and 3 until numerical convergence of the estimates

Output: Threshold variables, group structures, and a set of group-specific parameters.

In the following subsections, I explain the procedures in each step in more detail. First, I will describe step 2 of the algorithm. After, I will describe step 3. For the description of the two steps, it is convenient to first assume that the number of groups, G^0 is known. Later, in Section 3.4, I discuss the estimation of this number.

3.2 Estimating the threshold coefficient

The estimation of the threshold coefficient in step 2 of the algorithm is based on research conducted by Hansen (1999). The estimation uses the vector containing the threshold variable, q, to make a "selection" of threshold coefficients to consider. Ultimately, the procedure selects the optimal threshold, based on obtaining a minimal SSE.

In contrast with, for example, the slope coefficients, the threshold coefficient is defined on a specific set of values. Because of the fact that the threshold is merely used as a tool to split the data into groups, the exact value of the coefficient is less important than the selection which it creates. Therefore, testing the values in q as possible values for γ is sufficient to estimate the threshold coefficient. As mentioned above, this testing of the possible values for γ is based on the SSE of the model given a value for γ. Given γ, We can estimate the slope and threshold-effect coefficients by Ordinary Least Squares (OLS) procedure,

$$\hat{\theta}(\gamma) = (\tilde{x}(\gamma)'\tilde{x}(\gamma))^{-1}\tilde{x}(\gamma)'\tilde{y}, \tag{8}$$

such that the vector of residual values can be estimated as follows

$$\hat{e}(\gamma) = \tilde{y} - \tilde{x}(\gamma)'\hat{\theta}(\gamma). \tag{9}$$

Accordingly, we can calculate the SSE as follows

$$SSE(\gamma) = \hat{e}(\gamma)'\hat{e}(\gamma). \tag{10}$$

Minimising this SSE value for the threshold coefficient gives the optimal value for γ

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} SSE(\gamma). \tag{11}$$

The problem with this minimisation of the $SSE(\gamma)$ is that, as N and T take larger values, the testing of all different values of the threshold variable vector becomes very numerically intensive. Therefore, it is appropriate to reduce the number of possibilities for γ as much as possible, without significantly affecting the outcome of the estimate. This reducing of the search space is done by considering quantiles of the threshold variable vector, instead of the entire vector. This reduces the search space for γ greatly, while keeping it large enough to be sufficiently precise in most areas of interest. For the purpose of this research, I choose (in accordance with Hansen (1999)) a quantile grid of {1.00%, 1.25%, 1.50%, 1.75%, 2.00%, ..., 99.0%}. This reduces the search space for γ to 393 distinct values.

In terms of the algorithm, the estimation of group-specific threshold coefficients is required. The algorithm, therefore, estimates the value of the threshold coefficient for each group separately. This now leaves us with the estimation of the group-membership of individuals and the estimation of coefficients θ.

3.3 Estimation of group-membership, slope, and threshold-effect coefficients

The third step in the estimation algorithm calculates estimates for group-membership, g_i, slope, β_{g_i}, and threshold-effect, δ_{g_i}. This step is closely related to the C-Lasso technique proposed by

Su et al. (2016), as discussed in Section 2. This step of the algorithm constructs a Penalised Profile Likelihood (PPL) criterion function such that it is able to incorporate the threshold, and threshold-effect coefficients.

The profile log-likelihood function we consider in this C-Lasso approach is

$$Q_{1,NT}(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \tilde{x}'_{it}(\gamma)\theta_i \right)^2 \tag{12}$$

In order to allow the estimates for θ_i to shrink to the unknown group coefficients, α_g, we need to consider, in addition to the log-likelihood function, the minimisation of a penalty term. This error term, following Su et al. (2016), takes N additive terms, each of which includes the product of G^0 separate penalties. The resulting PPL criterion function to be minimised is

$$Q_{1NT,\lambda_1}(\theta, \alpha) = Q_{1,NT}(\theta) + \frac{\lambda_1}{N} \sum_{i=1}^{N} \prod_{g=1}^{G^0} \|\theta_i - \alpha_g\|. \tag{13}$$

The penalty term of this equation takes this multiplicative form, such that the criterion function allows each θ_i to shrink to any of the group-specific coefficients α_g. A priori, the group membership of θ_i is unknown, we cannot assign each individual-specific coefficient to a certain group-specific coefficient before the estimation process. The inclusion of this multiplication of the shrinkage effects of each of the α_g coefficients, permits θ_i to shrink to a particular group coefficient. The summation of the penalty of each of the N individuals then allows for the inclusion of information on the entire cross-section in identifying each β_i, $i = 1, ...N$ and α_g, $g = 1, ..., G^0$. The structure of this penalty, therefore, allows for the classification of each unknown β_i into unknown groups with unknown parameters, α_g. Furthermore, the penalty term is multiplied by a tuning parameter λ_1, which controls the strength of the penalty term.

3.4 Choice of the number of groups

Until now, we have assumed that the number of groups present in the data is known. However, in practice this is rarely the case. The number of groups, G, is then estimated, based on some criterion. This research uses an information criterion (IC) which, if minimised, gives the optimal value for G. Following Bonhomme and Manresa (2015), Su et al. (2016), and Miao et al. (2020), we consider an IC which takes into account both the error in the model and an additional term for the size of G. By adding a term for the number of groups in the model into the IC function to be minimised, we penalise larger values for G, keeping the model more straightforward. Assuming that the true number of groups, G^0, is bounded above by some value G_{max}, the IC considers possible values for G and tuning parameter, λ_1, and chooses the number of groups which minimises the following criterion function.

$$IC(G, \lambda_1) = \frac{2}{NT} \sum_{k=1}^{K} \sum_{i \in \hat{G}_k(G,\lambda_1)} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \tilde{x}'_{it}(\gamma_{g_i})\theta_{g_i} \right) + \rho_{1NT} pG \tag{14}$$

Where the penalty for big G, $\rho_{1NT}pG$, depends on the size of G, the number of variables, p, and some tuning parameter, ρ_{1NT}. We choose G as follows

$$\hat{G} = \underset{G \in \{1,\ldots,G_{max}\}}{\arg\min} \; IC(G, \lambda_1) \tag{15}$$

In the estimation algorithm, step 3 (the C-Lasso step) incorporates the re-estimation of optimal number of groups until convergence. As soon as the number of groups converges, this estimate, $\hat{G}$, can be set as the final number of groups and the algorithm can then skip the re-estimation of G.

4 Monte Carlo simulations

In order to evaluate the performance of the proposed estimation procedure in finite samples, I use Monte Carlo simulations. These simulations show how well the C-Lasso based PSTR model performs on simulated data, based on various fit measures. In this section, first I describe the data generating process (DGP) used in the simulations (Section 4.1). After, I propose some fit measures that evaluate the performance of the model (Section 4.2).

4.1 The data generating process

The DGP of the simulations concerns a static panel model. The data which the DGP generates contains distinct group-specific threshold influences, as well as group-specific influences of all variables. The data I generate comes from the following static panel model

$$y_{it} = \mu_i + x'_{it}\beta_{1,g_i} \cdot \mathbf{1}\{q_{it} \leq \gamma_{g_i}\} + x'_{it}\beta_{2,g_i} \cdot \mathbf{1}\{q_{it} > \gamma_{g_i}\} + \epsilon_{it}, \tag{16}$$

where I specify the values and distributions of the variables and parameters.

The sizes of the cross-section and time dimension vary over the different simulations. The values that N and T take are $50, 100$ and $30, 60$, respectively. Furthermore, G^0 is set to three. We can then split the cross-section dimension into these $G^0 = 3$ groups, and assign separate coefficients to each group. The DGP splits the data into groups by drawing data from three groups. The proportions of the groups are set to $N_1 : N_2 : N_3 = 0.3 : 0.3 : 0.4$.

The amount of variables, p, is two. In other words, the slope coefficients β_{1,g_i}, and β_{2,g_i} are 3×1 vectors of coefficients containing a group pattern of heterogeneity. Following the research by Miao et al. (2020), the slope parameters are set to

$$(\beta_{1,1}, \beta_{1,2}, \beta_{1,3})' = (1, 1.75, 2.5)', \quad \text{and} \quad (\beta_{2,1}, \beta_{2,2}, \beta_{2,3})' = (1, 1.75, 2.5)' + (NT)^{-0.1}$$

Where the addition of $(NT)^{-0.1}$ to the values of β_1 replicates the concept of a threshold-effect coefficient. Furthermore, the threshold variables for each group are defined by

$$(\gamma_1, \gamma_2, \gamma_3)' = (0.5, 1, 1.5)'$$

As for the distributions of the variables. The DGP draws observations from the following distributions

$$x_{it} \sim N(0, 1), \quad \text{and} \quad q_{it} \sim N(1, 1), \quad \text{and} \quad e_{it} \sim N(0, 1)$$

Where each of the variables are independent and identically distributed. The value for ϵ_{it} can be calculated from e_{it} by $\epsilon_{it} = (0.5 + 0.1x_{it}^2)^{1/2} \cdot e_{it}$.

For the simulations, I consider a set of 100 replications for estimation, which can be used to evaluate the performance of the model.

4.2 Fit measures of the estimates

The evaluation of the model calls for the interpretation of several statistics that can be examined following the simulation. In this subsection, I discuss the various fit measures that are used to investigate the performance of the model.

The first statistic which I examine is the frequency that the estimated value for G equals its true value. For this measure, the algorithm identifies the estimated number of groups and calculates the selection frequency of each value $G \in \{1, ..., G_{max} = 5\}$ by taking the total number of times each number of groups is selected and dividing this by the total number of replications.

To identify whether the algorithm can classify group membership of each individual accurately, I also look at the misclassification rate (MR) of the results. This misclassification rate is defined as the average number of misclassified individuals. I consider the following misclassification function per simulation iteration

$$MR = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{g}_i \neq g_i^0) \tag{17}$$

This MR value per simulation iteration is averaged over the number of replications to obtain a final misclassification rate of the estimation algorithm.

In addition to these measures describing the estimation of the groups, I use several fit measures to evaluate the estimated coefficient values. The bias and Root Mean Squared Error (RMSE) of each of the coefficients are measures that show how close the estimated coefficients are to the true values. Furthermore, the Coverage Probability (CP) of the 95% confidence level of the coefficients gives us an understanding of the accuracy of the coefficients.

For the slope, and threshold-effect coefficients, the CP can be calculated by using the normal confidence interval calculations. However, for the threshold coefficient, we do not have the standard deviation value needed to identify the confidence interval, since this coefficient is estimated from a set. The CP for this coefficient can be identified by considering a Likelihood Ratio (LR) test with $H_0 : \hat{\gamma}_g = \gamma_g^0$. Following the investigation by Hansen (1999), the LR test statistic and its distribution are defined as follows

$$LR(\gamma) = \frac{SSE(\gamma^0) - SSE(\hat{\gamma})}{\hat{\sigma}^2} \sim \zeta \tag{18}$$

Where $\hat{\sigma}^2$ is defined as $\hat{\sigma}^2 = \frac{1}{N_g(T-1)} SSE(\hat{\gamma}_g)$, where N_g is the number of individuals in group g. Furthermore, ζ is a random variable with corresponding distribution function $P(\zeta \leq x) = (1 - exp(-x/2))^2$, which has inverse $c(\alpha) = -2log(1 - \sqrt{1-\alpha})$. From this inverse, we calculate the critical value of the 95% confidence interval to be 7.35.

This LR test statistic allows us to calculate the CP for the threshold coefficient as well as for the slope, and threshold-effect coefficients.

5 Simulation results

In this section, I present the results obtained from the simulations as described in Section 4 using the C-Lasso based estimation algorithm. I present and discuss the performance of the algorithm with respect to the estimation of group-membership and the estimation of the actual coefficients.

5.1 Performance of group-structure estimation

As discussed in Section 4.2, I evaluate how well the algorithm estimates the group-memberships using two measures. First, I look at the selection frequency of the number of groups. After, I investigate the misclassification rate of the algorithm. These two measures give an idea of how well the algorithm assigns groups to each individual.

Table 1, shows the selection frequency of the number of groups $G \in \{1, ..., G_{max}\}$, where G_{max} is set to five.

Table 1: Selection frequency of number of groups, G

N	T	Number of groups				
		1	2	3	4	5
50	30	0.00	0.00	**0.98**	0.02	0.00
50	60	0.00	0.00	**1.00**	0.00	0.00
100	30	0.00	0.00	**0.97**	0.03	0.00
100	60	0.00	0.00	**1.00**	0.00	0.00

As we can see in the table above, a time dimension of only 30 leads to a probability of 97% and 98% of estimating the correct number of groups for cross-sections 50 and 100, respectively. In addition, doubling this dimension and estimating a model with $T = 60$, for both cross-section dimensions tested, gives us a probability of 100% for identifying the true value, G^0.

These results show that the proposed IC function, as described in Section 3.4, can identify the number of groups present in the data set with a probability increasing up to 100% as the size of the time dimension of the panel data is increased. This indicates that the this estimation of number of groups is very accurate and suitable. This ability of the algorithm to correctly specify the value of G every time, leaves the evaluation of the actual group-allocation of individuals, the measure for which is the misclassification rate.

The misclassification rates for the method are given in Table 2.

Table 2: Misclassification rates

N	T	MR	N	T	MR
50	30	0.0433	100	30	0.0410
50	60	0.0075	100	60	0.0064

The misclassification rates (MRs) we can see here, similarly to the selection frequency results discussed above, show promising results. For the simulations with time dimension $T = 30$, the MRs are already fairly low. The rates, for both cross-section dimensions, of around 0.04 indicate that, on average, 4% of the individuals are misclassified into the wrong group. However, doubling the time dimension decreases this percentages drastically. For a time dimension of $T = 60$, over 99% of individuals are allocated to their true group, g_i^0.

All in all, from these results, we can conclude that the estimation algorithm is very accurate in estimating the group structures present in the data. This allows us to evaluate the performance of the algorithm with regard to the consistent estimation of the coefficients, given that each individual is correctly classified into their group.

5.2 Performance of coefficient estimation

Now that I have investigated the ability of the algorithm to allocate the correct group to each individual, I evaluate the group-specific slope, threshold-effect, and threshold coefficients based on their bias, RMSE value and coverage probability. All results are presented in Table 3.

Table 3: Estimates of coefficients and threshold values

		β_1			β_2			γ		
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	RMSE	CP
N=50	Group 1	0.002	0.095	0.89	0.001	0.057	0.89	0.027	0.150	0.86
T=30	Group 2	-0.016	0.108	0.78	0.035	0.081	0.82	0.034	0.152	0.95
	Group 3	0.009	0.058	0.86	0.014	0.080	0.87	0.024	0.113	0.91
N=50	Group 1	0.004	0.054	0.93	-0.003	0.035	0.96	-0.008	0.063	0.91
T=60	Group 2	-0.005	0.050	0.93	0.007	0.047	0.91	-0.004	0.054	0.91
	Group 3	0.001	0.035	0.91	0.009	0.039	0.98	0.012	0.062	0.91
N=100	Group 1	-0.001	0.051	0.95	0.016	0.043	0.90	-0.005	0.088	0.88
T=30	Group 2	-0.009	0.066	0.85	0.041	0.065	0.83	0.001	0.069	0.95
	Group 3	0.014	0.042	0.83	0.013	0.063	0.90	-0.005	0.052	0.95
N=100	Group 1	-0.009	0.040	0.95	-0.002	0.020	0.98	0.003	0.037	0.95
T=60	Group 2	-0.009	0.030	0.95	0.009	0.032	0.93	0.001	0.052	0.93
	Group 3	0.000	0.025	0.93	0.008	0.030	0.95	-0.003	0.025	0.95

Looking at the estimation results for the coefficients for $T = 30$, we can see that most CP values lie between 85% and 95%. For the simulation with the smaller cross-section, the probabilities are slightly lower than for the simulation with $N = 100$. We can see a similar trend in bias and RMSE values for these simulations. Overall, while still being small values, both the bias and RMSE values for the simulation with $N = 50$, $T = 30$ are slightly higher than those for the simulation with $N = 100$, $T = 30$. However, the results for both simulations show promising bias, RMSE, and CP values.

The simulations with time dimension $T = 60$ are even more promising. Looking at both the simulation with $N = 50$, and $N = 100$, we see the same trend as for the smaller time dimension. The results are slightly better when the cross-section dimension is greater. Overall, we can conclude that extending the time-dimension in the panel data set greatly increases the performance of the estimation algorithm. Reaching CP values of up to 0.98, and bias values as low as 0.000 allows for the conclusion that, especially for a big time dimension, the estimation algorithm can very accurately estimate all group-specific coefficient values in the model.

6 Empirical application

In this section, in order to show the applicability of the estimation method, a apply the algorithm proposed in this research to an empirical data set. The data set I use in this section is an extension of the one used in Su et al. (2016). The set considers data on savings rates of various countries. In literature, understanding the savings behaviour of countries has been greatly investigated (see for example Su et al. (2016) and Bosworth et al. (1999)). Research into this area of development economics is generally based on standard panel structure models which, in some cases, incorporates the heterogeneity in the data using various techniques. In Su et al. (2016), for example, the researchers use their C-Lasso method to take into account the latent group structures in the data. In addition, Loayza et al. (2000), for example, use simple classification criteria to group the countries. However, in past literature, PSTR models with group-specific coefficients has not been applied to the data under consideration.

The data I use concerns a set taken from the World Development Indicators from the World Bank data base. The set, as used by Su et al. (2016), includes data on savings rates, inflation rates, interest rates and GDP of 56 countries over a period of 15 years. The set considers yearly data points between 1996 and 2010. The researchers, in their investigation, show the presence and estimation of latent group structures in the data. Following research by Fouquau and Hurlin (2008), I include a variable concerning the current account balance of the countries as a threshold variable.

The addition of this variable to the data set as used in Su et al. (2016) leads to an unbalanced panel set. For simplicity, I decide to remove the incomplete observations. This leaves a data set containing 55 countries over a period of 15 years. An overview of some information concerning the various variables can be found in Table 5 in Appendix A.1. This table contains some descriptive statistics of all variables in the model. Furthermore, Appendix A.2 contains a list of all countries that are in the data set.

The model for the savings rate of each country is

$$S_{it} = \mu_i + (I_{it} + R_{it} + G_{it})\beta_i + (I_{it} + R_{it} + G_{it})\delta_i \cdot \mathbf{1}\{C_{it} \geq \gamma_i\} + \epsilon_{it}, \tag{19}$$

with S_{it} containing the savings rate of each country per year. I_{it} is the inflation rate, R_{it} the real interest rate, and G_{it} the GDP growth per country per year. Furthermore, C_{it} contains the current account balance data. The variables I include in the model are similar to the ones used by Su et al. (2016), apart from the current account balance which I add as threshold variable. I, in accordance with Su et al. (2016), add the inflation and interest rate as independent variables to identify the value of money and degree of economic stability of each country. Furthermore, I add the GDP growth as a variable since research (see Saltz (1999)) has shown a correlation between GDP and the savings of countries. I use the current account balance as threshold variable to include a measure of trade in goods and services into the model.

The estimation algorithm, based on the IC function described in Section 3.4, chooses the number of groups in the data to be 2, and value for the tuning parameter, λ_1, to be 0.7188. Figure 1 shows allocation of group-memberships, estimated by the algorithm, of the countries in the data set. It is interesting to see that most Asian countries fall into group 2, while more Western countries fall in group 1. This shows that geographical placing of countries, most likely, has some effect on the results.

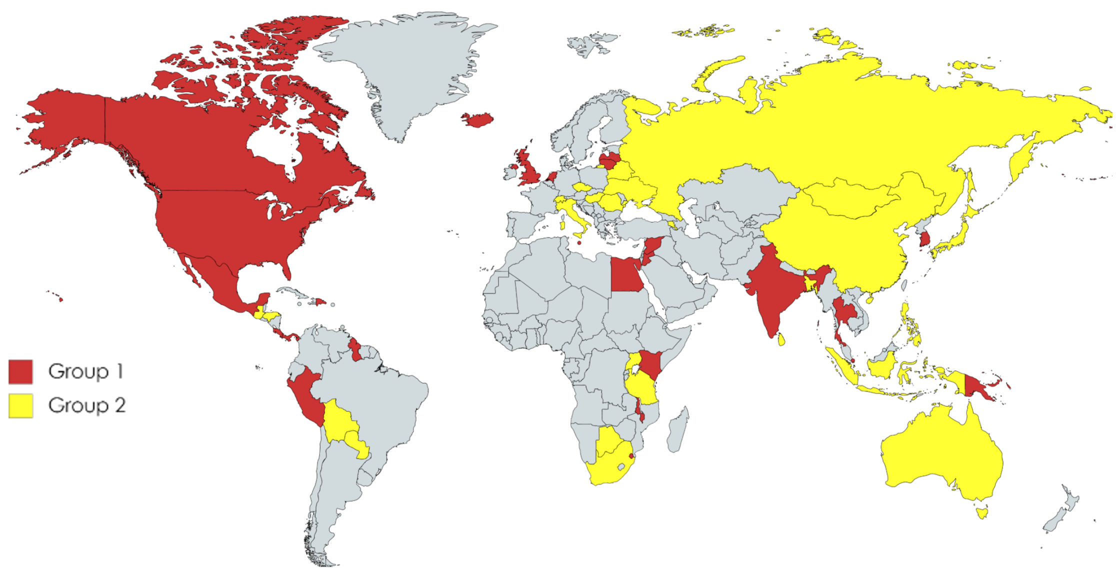

Figure 1: Group-membership of countries in the data set

The estimated coefficients can be found in Table 4. We should note that, since the threshold coefficient is estimated from a set, it does not have a standard deviation. Furthermore, its significance is evaluated by testing the hypothesis $H_0 : \hat{\gamma}_g = 0$, and investigating the significance level for which we can reject this null hypothesis.

This specification of current account balance as a threshold variable for the savings rates of countries splits the data into a group above and a group below the threshold. In group 1, 10.2% of the data falls below the threshold and in group 2, 33.3% falls below. Both groups have significant threshold coefficients which isolates a small part of the data set to be below the threshold. This shows that the effect the independent variables have on the savings rate is different for countries with a measure of trade which is very low compared to countries with a relatively higher account balance. For group 1, the threshold coefficient isolates the countries with the lowest 10% of relative current account balance, while for group 2 this value is slightly higher. Both values, however, indicate that mostly the model can be estimated with the same model and that only a small percentage of observations have independent variables with a significantly different effect on the dependent variable.

Table 4: Savings rate data: estimation results

Variables		Group 1	Group 2
	Inflation	-0.1396 (0.1594)	-0.3373 (0.0850)
β	Interest rate	0.3806* (0.1909)	-0.6285 (0.0923)
	GDP growth	0.2305** (0.1079)	0.2882*** (0.0790)
	Inflation	0.5474*** (0.1694)	0.1933** (0.0989)
δ	Interest rate	-0.1605 (0.1984)	0.2820*** (0.1039)
	GDP growth	-0.1366 (0.1193)	0.3611*** (0.0973)
γ	Current account balance	-8.7031***	-4.4574***

Note: *10% significant, **5% significant, ***1% significant.
Standard deviations between brackets.

Looking at the estimated coefficient values, we can see for both groups, a 1% significant positive GDP growth coefficient value. This supports the conventional views and previous research on the correlation between GDP and savings rate (see for example Saltz (1999)). This effect increases significantly if the current account balance lies above the threshold for group 2. As for the inflation and interest rate, we can see that the β coefficients are insignificant, apart from the 10% significance of the interest rate for group 1. In accordance with Edwards (1996) and Su et al. (2016), this result reiterates the importance of the effect of GDP growth on savings rates.

7 Conclusion

This paper investigates the performance of a C-Lasso based estimation for PSTR models with latent group structures. I aim to develop a Lasso based technique which can accurately estimate all group-specific coefficients in PSTR models without restricting the freedom of the model. This innovative method uses an algorithm with several steps. One step updates the threshold coefficient values for each of the groups, while the other step estimates the number of groups, group-membership and slope, and threshold-effect coefficient values using a C-Lasso type estimation technique. The performance of the model is evaluated by means of a Monte Carlo simulation which allows for the interpretation of various fit measures of the estimates, as well as the evaluation of the ability to correctly identify the number of groups and group-memberships. Lastly, I use a panel data set containing information of savings rates of 55 countries to show the empirical application of this new method.

The simulation results show that the proposed algorithm performs well in finite samples. The estimation of group-membership results in a very high probability of choosing the correct number of groups, as well as correctly allocating group-memberships. Furthermore, I show that the algorithm can quite accurately estimate the group-specific coefficients. The values for bias, as well as the RMSE values showed promising results, revealing that the estimated values for the coefficients lie close to the true values of the coefficients. Furthermore, the coverage probability of the coefficients show the probability of correctly estimating the coefficients in a 95% confidence interval. These probability values were around 90% or even higher for all estimates.

Another important finding from the simulation study is the great influence of the size of the time dimension on the estimation results. As T increases, the estimation of true number of groups approaches a probability of 100%. In addition, the misclassification rates for such T take very small values. The same trends of increasing performance are also found in the fit measures for the estimated coefficients.

From these results, I conclude that the algorithm accurately estimates a PSTR model with latent group structures. Furthermore, the empirical example shows the applicability of the algorithm and reiterates the correlation between GDP and savings rates in countries.

There are some interesting topics for possible future investigation. Firstly, in this research, I only consider linear panel data models. Research into, for example, non-linear models could improve the applicability of the model. Furthermore, investigation into the possibility of endogenous regressors and threshold variables could lead to an even better understanding of PSTR models. Lastly, one could also investigate allowing for time effects instead of, or in addition to, cross-sectional fixed effects.

References

Ando, T. and Bai, J. (2015). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bosworth, B. P., Collins, S. M., and Reinhart, C. M. (1999). Capital flows to developing economies: Implications for saving and investment. *Brookings Papers on Economic Activity*, 1999(1):143.

Caner, M. and Hansen, B. E. (2004). Instrumental variable estimation of a threshold model. *Econometric Theory*, 20(05).

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 21(1):520–533.

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014). Group lasso for structural break time series. *Journal of the American Statistical Association*, 109(506):590–599.

Edwards, S. (1996). Why are latin americas savings rates so low? an international comparative analysis. *Journal of Development Economics*, 51(1):5–44.

Fouquau, J. and Hurlin, C. (2008). The feldstein-horioka puzzle: A panel smooth transition regression approach. *SSRN Electronic Journal*.

Girma, S. (2005). Absorptive capacity and productivity spillovers from fdi: A threshold regression analysis. *Oxford Bulletin of Economics and Statistics*, 67(3):281–306.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics*, 93(2):345–368.

Huang, B.-N., Hwang, M., and Peng, H.-P. (2005). The asymmetry of the impact of oil price shocks on economic activities: An application of the multivariate threshold model. *Energy Economics*, 27(3):455–476.

Kremer, S., Bick, A., and Nautz, D. (2012). Inflation and growth: new evidence from a dynamic panel threshold analysis. *Empirical Economics*, 44(2):861–878.

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1).

Loayza, N., Schmidt-Hebbel, K., and Serven, L. (2000). Saving in developing countries: An overview. *The World Bank Economic Review*, 14(3):393–414.

McLachlan, G. J. and Peel, D. (2006). *Finite mixture models*. John Wiley amp; Sons.

Miao, K., Su, L., and Wang, W. (2020). Panel threshold regressions with latent group structures. *Journal of Econometrics*, 214(2):451–481.

Poetscher, B. M. and Leeb, H. (2007). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *SSRN Electronic Journal*.

Saltz, I. S. (1999). An examination of the causal relationship between savings and growth in the third world. *Journal of Economics and Finance*, 23(1):90–98.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, 21(1):3.

Su, L., Shi, Z., and Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

A Additional information for empirical application

A.1 Descriptive statistics

Table 5: Descriptive statistics

Variable	Mean	Std. Dev.	Min	Max
Savings rate	22.089	8.889	-3.207	53.434
Inflation	7.793	15.458	-2.478	293.679
Interest rate	7.517	10.010	-63.761	93.915
GDP growth	2.879	3.879	-17.545	14.060
Current account	-1.606	6.899	-24.229	27.143

A.2 List of countries in the cross-section

The countries in the data set used for the empirical application in this research are (in alphabetical order):

Armenia, Australia, The Bahamas, Bangladesh, Belarus, Bolivia, Botswana, Canada, Cape Verde, China, Costa Rica, Czech Republic, Dominican Republic, Egypt, Guatemala, Guyana, Honduras, Hungary, Iceland, India, Indonesia, Israel, Italy, Japan, Jordan, Kenya, South Korea, Latvia, Lithuania, Malawi, Malaysia, Malta, Mauritius, Mexico, Mongolia, The Netherlands, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Romania, Russia, Singapore, South Africa, Sri Lanka, Swaziland, Switzerland, Tanzania, Thailand, Uganda, Ukraine, United Kingdom, United States, Urugay.