Panel Data Analysis Via Variable Selection and Subject Clustering

Haibing Lu, Shengsheng Huang, Yingjiu Li and Yanjiang Yang

Abstract A panel data set contains observations on multiple phenomena observed over multiple time periods for the same subjects (e.g., firms or individuals). Panel data sets frequently appeared in the study of Marketing, Economics, and many other social sciences. An important panel data analysis task is to analyze and predict a variable of interest. As in social sciences, the number of collected data records for each subject is usually not large enough to support accurate and reliable data analysis, a common solution is to pool all subjects together and then run a linear regression method in attempt to discover the underlying relationship between the variable of interest and other observed variables. However, this method suffers from two limitations. First, subjects might not be poolable due to their heterogeneous nature. Second, not all variables might have significant relationships to the variable of interest. A regression on many irrelevant regressors will lead to wrong predictions. To address these two issues, we propose a novel approach, called *Selecting and Clustering*, which derives underlying linear models by first selecting variables highly correlated to the variable of interest and then clustering subjects into homogenous groups of the same linear models with respect to those variables. Furthermore, we build an optimization model to formulate this problem, the solution of which enables one to select variables and clustering subjects simultaneously. Due to the combinatorial nature of the problem,

H. Lu (✉)
Santa Clara University, Santa Clara, United States
e-mail: hlu@scu.edu

S. Huang
University of Houston—Victoria, Victoria, United States
e-mail: huangs@uhv.edu

Y. Li
Singapore Management University, Singapore, Singapore
e-mail: yjli@smu.edu.sg

Y. Yang
Institute for Infocomm Research, Singapore, Singapore
e-mail: yyang@i2r.a-star.edu.sg

K. Yada (ed.), *Data Mining for Service*, Studies in Big Data 3,
DOI: 10.1007/978-3-642-45252-9_5,

an effective and efficient algorithm is proposed. Studies on real data sets validate the effectiveness of our approach as our approach performs significantly better than other existing approaches.

1 Introduction

Panel data can be defined as the data set with the structure consisting of different subjects (e.g., countries, states, patients) with multiple observations (e.g., at an annual, quarterly, monthly, or hour base) in a certain time period. A benefit of the panel data is that they provide two dimensions of variation–the cross-sectional and time series variations to trace the change of subjects over time and at the same time to overcome the problem of limited observations per subject by pooling different subjects together. Panel data analysis is widely used in social sciences such as economics, finance and marketing science [1].

One of the most important panel data analysis tasks is to study the relationship between the variable of interest and other observed variables by examining data observations. The relationship can further be employed to make predictions on the variable of interest. One challenge is that for many real panel data sets encountered in social sciences, the number of observations for each subject is often small with respect to the number of observed variables. As a result, if one tries to derive the variables relationship for each individual subject, it is unlikely to obtain an accurate and reliable estimation. To illustrate, look at subjects 1–3 in Fig. 1. As each subject has only four observations, there is no significant relationship between the dependent variable and the independent variable that can be observed. To address this issue, a common approach in the literature is to pool all subjects together and run a linear regression method to estimate the linear relationship between the variable of interest and the other observed variables. Consider the same example. If we pool subjects 1–3 together as illustrated in Fig. 1d, a significant linear relationship between the dependent and independent variables might surface.

However, the success of this pooling approach is based on two assumptions. The first assumption is that all subjects are homogenous, which means observations of every subject are generated from the same model. The second assumption is that the variable of interest is dependent on all observed variables. In other words, the variable of interest is strongly correlated with all observed variables. Unfortunately, in many real cases, these two assumptions do not hold.

Consider the example of Fig. 2. In Fig. 2c, there are five subjects including the three subjects of Fig. 1a, b, c and the two subjects of Fig. 2a, b. There are obviously two linear relationships. In other words, those five subjects are partially homogenous such that subjects 1–3 belong to one homogenous group and subjects 4–5 belong to the other homogenous group. In this case, the first assumption that all subjects are homogenous clearly does not hold. If one has to pool all of the five subjects together, a wrong conclusion on the relationship between variables will be made.

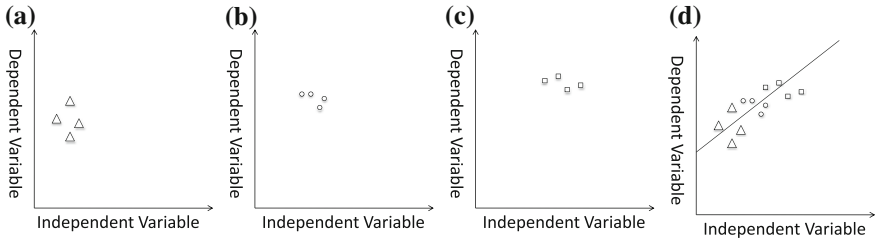

Fig. 1 An example of homogenous subjects. **a** Subject 1, **b** Subject 2, **c** Subject 3, **d** Collection

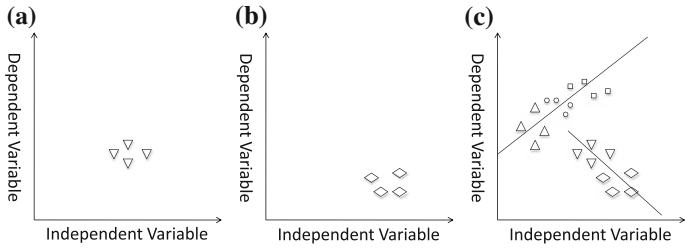

Fig. 2 An example of partially homogenous subjects. **a** Subject 4, **b** Subject 5, **c** Collection

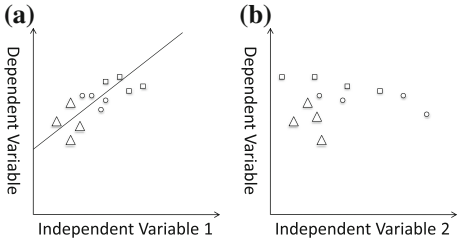

Fig. 3 Illustration of necessity of variable selection. **a** with independent variable 1, **b** with independent variable 2

Next consider the example of Fig. 3. Different subjects are plotted with different shapes. Each subject has been characterized with three variables, the dependent variable and the independent variables 1 and 2, where the dependent variable is the variable of interest. A significant linear relationship between the dependent variable and the independent variable 1 can be observed in Fig. 3a. However, the independent variable 2 seems uncorrelated with the dependent variable as shown in Fig. 3b. As a result, if we need to find the common pattern within the subjects using the three observed characteristics (i.e., three variables), inclusion of variable 2 might blurs the existing pattern captured by the dependent variable and the variable 1. Therefore, in the exploratory studies in search for potential patterns or theories in the data, variable selection is critical. Especially when a strong theory guiding the regression model is lacking, a valid and efficient approach for variable selection is valuable. This toy example illustrates the necessity of variable selection in panel data analysis.

To address those two issues, we propose to discover underlying linear models through two main steps: (1) identifying variables significantly related to the variable of interest, (2) and based on those variables clustering multiple subjects into homogeneous groups such that each group belongs to the same model. The approach is called *Selecting and Clustering*. The BIC information criteria is used to help determine the number of clusters and the number of variables selected to be included in the linear models. Needless to say, as a general rule, the final solution of variables selected and clusters identified depends on the information criteria applied by researchers. However, whatever the information criterion used, the method in this study and its contribution will not change. An optimization model is built to formulate this model selection problem with BIC as the information criteria. Its solution selects important variables and clusters subjects simultaneously. Due to the combinatorial nature of this optimization model, we propose a simulated annealing based search strategy to traverse possible variable selection solutions and also present an iterative algorithm to perform the clusterwise linear regression to cluster subjects and discover underlying linear models.

The remainder of the chapter is organized as follows. Section 2 reviews related work. Section 3 formally introduces the variable selection and subject clustering problem. Section 4 presents an effective heuristic to deal with the presented problem. Section 5 conducts experimental studies on the statewide productivity data and the OECD gasoline demand data and Sect. 6 concludes the chapter.

2 Related Work

While panel datasets can enable more accurate analysis for complex phenomena, an important issue in analyzing panel data is the poolability of different subjects [2]. In a linear regression model that is the typical statistic approach in social sciences, the different subjects can be pooled together if the parameters in the regression (i.e., coefficients of independent variables and constant term) can be considered homogenous across different subjects. Some of the tests, such as Chow test and Wald test (e.g., [3]) can be extended to check the poolability before analyzing panel data [4]. If the dataset fails the poolability test, the pooled panel data would not produce the estimates that are both theoretically and statistically valid, due to the heterogeneity between subjects implied by significant difference of the parameters. In conventional practices, scholars either abandoned the pooling approach to estimate the model coefficients [5], or used a weighting method to "shrink" the individual estimates toward the pooled estimate [6]. However, ignoring the partial commonality shared by the subjects (i.e., assuming full heterogeneity) could lead to very imprecise parameter estimates [7], due to limited observations per subject.

More recent development in econometrics is the investigation into "partial poolability" of the data and tried to cluster the subjects into different groups so that the subjects within a same group (or cluster) are homogenous in terms of the effect of independent variables. For example, Vahid [7] adopts the likelihood ratio statistic

testing the equality of the parameters between two subjects as a distance measure. Then clustering was based on this distance measure and an additional sub-group consistency measure. Kapetanios [4] proposed an intuitive method to determine the cluster membership by comparing the information criteria statistics from all possible clustering solutions. Heuristic algorithms are used to overcome the overwhelming computation burden incurred by this method, such as Simulated Annealing [4] and Expecting-Maximization [8].

The extension from poolability test to clustering the panel data has significant implications for model building and theory development in economics and finances. The different effect size is exactly the main argument of the research stream exploring heterogeneity between subjects. For example, evolutionary economics emphasizes the different pattern of economic growth of countries (e.g., [3, 9–12]). Most of the existing studies used pre-selected variables as the criteria for clustering (e.g., CART approach [3]). Obviously, different grouping variables could lead to different and even conflicting results. Moreover, in many cases, it is very difficult to choose a legitimate grouping variable, or to decide the critical values of that variable to distinguish groups. In some cases, there is very little a-priori knowledge (or reason) about which variables could be used. Accordingly, clustering panel data based on the true effect size from the data per se provides a useful method to minimize above limitations.

In the language of data mining, clustering panel data can be viewed as prototype-based clustering. All existing panel clustering approaches in economics and finance assume that the prototype of each cluster is a linear function and determine cluster memberships by the fitness of the data. Similar topics have been studied in data mining under different names, for example, *Regression Clustering* [13], *Clusterwise Linear Regression* [8, 14], *Trajectory Clustering Using Mixtures of Regression Models* [15], and *Clustered Partial Linear Regression* [16]. Unfortunately, they suffer from the same limitations.

3 Methodology

We consider the problem that: (1) a panel data set contains observations of multiple subjects with multiple variables, (2) there are linear relationships between the variable of interest and some observed variables, (3) each subject belongs to one linear model, and (4) the concrete variables correlated with the variable of interest, the number of linear models and their coefficients are unknown. The goal is to identify the variables that are correlated with the variable of interest, discover the underlying linear models, and determine the model memberships.

We tackle this problem by formulating it into an optimization problem. Our optimization model is facilitated by the following definitions:

- I = the number of subjects, indexed by $i = 1, ..., I$;
- J = the number of independent variables, indexed by $j = 1, ..., J$;

- T = the number of observation periods, indexed by $t = 1, ..., T$;
- K = the number of segments, indexed by $k = 1, ..., K$;
- L = the number of independent variable correlated with the the dependent variable, indexed by $r_l = r_1, ..., r_L$;
- V = a $J \times 1$ binary vector with $v_j = 1$ if variable j is correlated with the dependent variable, otherwise $v_j = 0$;
- X = an $I \times J \times T$ matrix with elements x_{ijt} representing the measurement value of subject i on independent variable j at period t for all $i = 1, ..., I$, $j = 1, ..., J$ and $t = 1, ..., T$;
- Y = an $I \times t$ matrix with elements y_{it} representing the measurement value of subject i on the dependent variable at period t for all $i = 1, ..., I$ and $t = 1, ...T$;
- P = an $I \times K$ binary matrix where $p_{ik} = 1$ if subject i is assigned to segment k and 0 otherwise, for all $i = 1, ..., I$ and $k = 1, ..., K$;
- α_k = the regression intercept value for cluster, for all $k = 1, ..., K$;
- β_{jk} = the regression equation slope coefficient for independent variable j in cluster k.

Given the above definitions, our problem can be formally restated as the following.

Problem 1. Given observations of J independent variables and one dependent variable for I subjects, discover the K underling linear models

$$y_k = \alpha_k + \sum_{r_l=r_1}^{r_L} \beta_{r_l k} x_{r_l k} + \varepsilon_k, 1 \leq k \leq K \tag{1}$$

where y_k is the response variable, $x_{r_l k}$ are explanatory variables, $\beta_{r_l k}$ are coefficients, α_k is intercept, and ε_k is the model errors.

In reality, the number of linear models K is usually unknown. A good model should balance goodness of fit with simplicity. Information criteria are often used for model selection. Bayesian information criterion (BIC) is one of the most popular information criteria used for linear model selection. Under the assumption that the model errors are independent and identically distributed according to a normal distribution, the BIC formula can be represented as the following:

$$BIC = I \times log(RSS) + T \times log(I) \tag{2}$$

where RSS is the residual sum of squares, I is the number of data points, and T is the number of free parameters to be estimated.

The model selection problem is then translated into an optimization problem: finding a set of linear models which minimizes the resultant BIC value.

So our methodology is to find a set of linear models which can fit the data well and also have a simple form. Mathematically, it is to solve the following optimization problem.

$$min\ I \times log(RSS) + K \times L \times log(I)$$

$$s.t. \begin{cases} RSS = \sum_k \sum_t \sum_{i|p_{ik}=1} (\alpha_k + \sum_{j=1}^{J} \beta_{jk} x_{ijt} - y_{it})^2 \\ L = \sum_k v_k \\ \beta_{jk} = 0 \ \ if\ v_k = 0\ \forall j \\ \sum_k p_{ik} = 1 \\ p_{ik} \in \{0, 1\} \\ v_k \in \{0, l\} \end{cases} \tag{3}$$

The detailed explanation to the above optimization model is given as the followings:

- The objective function is to minimize the BIC value of the linear models to be discovered;
- In the objective function, $K \times L$ is the total number of non-zero (effective) coefficients and evaluates the simplicity of the discovered linear models;
- RSS, which stands for residual squared sum, evaluates the goodness of fit and is computed as the first constraint;
- L is the number of independent variables correlated to the dependent variable.
- The constraint of $\beta_{jk} = 0$ if $v_k = 0\ \forall j$ ensures all linear models have the same explanatory variables.
- The constraint $\sum_k p_{ik} = 1$ ensures that each data point is assigned to one and only one cluster (linear model);
- Variables to be determined are: K the number of clusters, $\{\alpha_k, \beta_{jk}\}$ the coefficients of linear models, and p_{ik} the cluster assignments.

4 Algorithms

In this section, we will study how to solve the presented subject clustering and variable selection problem. The problem is essentially a combinatorial problem as each feasible solution is a combination of a partition of subjects and a selection of a variable set. To simplify the problem, we temporarily assume that the number of partitions, K, and the number of correlated variables, L, are known. If one can find an efficient algorithm to solve this simplified case, he/she can easily extend it to deal with the general problem by repeating the algorithm with different values of K and L and selecting the best one. As panel data sets encountered in social sciences

are usually not large, if the algorithm for the simplified has good performance, its generalized algorithm still works practically.

Now we focus our attention on the simplified problem that both of the number of partitions, K, and the number of correlated variables, L, are known. In this case, the complexity of linear models is fixed. Therefore, with respect to the BIC criteria, one only needs to minimize the residual sum of square.

In fact, each feasible solution is associated with a partition of subjects into K groups and a selection of L variables out of J variables. Both of the partition problem and the selection problem are NP-hard in general.

So we propose to tackle them by

- calling a simulated annealing algorithm to select L variables out of the J variables and
- calling an iterative algorithm to partition I subjects into K groups.

Note that both proposed solutions are heuristics. Ideally, if one traverses all possible combinations of subject partition and variable selection, the global optimum can be reached. However, it would be computationally expensive. The basic idea of our proposed solution is to apply a simulated annealing strategy to select a variable subset, which is a heuristic approach, and then given the selected variable subset, apply an iterative algorithm to partition subjects into groups and infer linear models.

4.1 Simulated Annealing for Variable Selection

We will present a simulated annealing algorithm for variable selection. Simulated annealing is a generic probabilistic heuristic for the global optimization problem. It locates a good approximation to the global optimum of a given function in a large search space. Different from greedy heuristics, simulated annealing is a generalization of a Markov Chain Monte Carlo method, which has a solid theoretical foundation.

A variable selection solution can be represented by a binary vector V of size $J \times 1$, where $v_i = 1$ means variable i is selected. Given large values of J and L, it is difficult to consider every possible combination of L variables out J variables. Simulated annealing is a strategy allowing one to spend less time and obtain a satisfactory solution.

We present the simulated annealing heuristic for the variable selection as the following. First, we let $(0, ..., 0)$ be the starting state of V. Then at each stage, randomly select its neighboring value by randomly picking one element of V and flipping its value from 0 to 1 or from 1 to 0. If the new V reduces fitting errors, the next state is the new V. If not, with a certain probability less than 1, the next state is still the new V. In other words, with certain probability, it remains its original state. This property reduces the chance of being stuck at a local optimum. The procedure described above allows a solution state to move to another solution state and hence produces a Markov Chain. Denote the nth state be V

Algorithm 1 Simulated Annealing Algorithm for Variable Selection

Input: $X, Y, K, L, limit$
Output: V
1: $V \leftarrow (0, ..., 0)$; $n \leftarrow 1$;
2: **while** $n \leq limit$ **do**
3: V'= a random neighboring value of V;
4: **if** $RSS(X, Y, K, V) < RSS(X, Y, K, V')$ **then**
5: $V \leftarrow V'$;
6: **else**
7: $V \leftarrow V'$ with probability $min\{1, \frac{exp\{log(1+n)\cdot RSS(X,Y,K,V)\}}{exp\{log(1+n)\cdot RSS(X,Y,K,V)\}}\}$;
8: **end if**
9: $n \leftarrow n + 1$;
10: **end while**

and the randomly selected neighboring value be V'. If the next state is V' with probability

$$min\left\{1, \frac{exp\{\lambda V(x)/N(x)\}}{exp\{\lambda V(y)/N(y)\}}\right\}$$

or it remains V, where λ is a constant, $V(t)$ is the reconstruction error with the solution t, and $N(t)$ is the number of neighboring values of t. Such a Markov Chain has a limiting probability of 1 for arriving at optimal minimization solutions when $\lambda \rightarrow \infty$ [17]. But it has been found to be more useful or efficient to allow the value of λ to change with time. Simulated annealing is a popular variation of the preceding. Here, we adopt the formula proposed by Besag et al. [18] and let the transition probability be

$$min\left\{1, \frac{exp\{\lambda_n V(x)/N(x)\}}{exp\{\lambda_n V(y)/N(y)\}}\right\}$$

where $\lambda_n = log(1+n)$. In our case, $N(x)$ and $N(y)$ are equivalent and are canceled out in the formula. As computing time is limited, we terminate the algorithm after a certain number of iterations regardless of whether or not the global optimum is reached.

Our complete simulated annealing algorithm is described as in Algorithm 1. The loop terminating condition is that the number of repetitions is less than a predefined number $limit$. The input of the function $RSS(X, Y, K, V)$ consists of the observations of independent variables X, the observations of dependent variables Y, the number of partitions K, and the selected variables V. The output of the function $RSS(X, Y, K, V)$ is the optimal residual sum of squares. How to compute $RSS(X, Y, K, V)$ will be studied in the Sect. 4.2.

4.2 Iterative Algorithm for Subject Clustering

In this section, we will study how to compute $RSS(X, Y, K, V)$, the optimal residual sum of squares, which is a clusterwise linear regression problem.

If K and V are given, it is not difficult to compute $RSS(X, Y, K, V)$ as it can be solved through the ordinary least squares method (OLS). As V is given after the variable selection step, then the difficulty lies in clustering subjects into k homogenous groups while minimizing $RSS(X, Y, K, V)$. It is a typical clusterwise linear regression problem, which is known to be NP-hard. To tackle it, we propose an iterative algorithm. Before getting to it, we first discuss the ordinary least square (OLS) method.

4.2.1 Ordinary Least Squares

The OLS method estimates the unknown parameters in a linear regression model by minimizing the sum of squared distances between the observed responses in the dataset, and the responses predicted by the linear approximation.

Suppose that we have a data set of $\{y_t, x_{t1}, ..., x_{tJ}\}_{t=1}^{T}$. Assume that the linear relationship between the dependent variable y_i and the vector of regressors x_i is linear. The linear model takes the form

$$y_t = \alpha + \beta_1 x_{t1} + ... + \beta_J x_{tJ} + \varepsilon_t = \alpha + \beta' X_t + \varepsilon_t, t = 1, ..., T \tag{4}$$

where X_t denotes the vector of $(x_{t1}, ..., x_{tJ})'$.

The OLS method is to estimate $\{\alpha, \beta\}$ by minimizing the residual sum of squares

$$f(\alpha, \beta) = \sum_t (\alpha + \beta' X_t - y_t)^2. \tag{5}$$

In other words,

$$(\alpha, \beta) = arg\ min\ \sum_t (\alpha + \beta' X_t - y_t)^2. \tag{6}$$

By denoting $\begin{pmatrix} \beta \\ \alpha \end{pmatrix}$ by $\mathscr{B}$, we rewrite the above formula as the following:

$$f(\mathscr{B}) = \sum_t ((X_t, 1)\mathscr{B} - y_t)^2. \tag{7}$$

Furthermore we put the above formula in a matrix form as below:

$$\begin{aligned} f(\mathscr{B}) &= (\mathscr{X}\mathscr{B} - Y)'(\mathscr{X}\mathscr{B} - Y) \\ &= \mathscr{B}'\mathscr{X}'\mathscr{X}\mathscr{B} - 2Y'\mathscr{X}\mathscr{B} + Y'Y \end{aligned} \tag{8}$$

where $\mathscr{X} = \begin{pmatrix} X_1', 1 \\ \ldots \\ X_T', 1 \end{pmatrix}$ and $Y = \begin{pmatrix} y_1 \\ \ldots \\ y_T \end{pmatrix}$.

Since this is a quadratic expression and $f(\mathscr{B}) \geq 0$, the global minimum can be found by differentiating it with respect to $\mathscr{B}$. The deduction steps are as the following:

$$\begin{aligned} \nabla f(\mathscr{B}) &= 0 \\ \mathscr{B}' \mathscr{X}' \mathscr{X} - Y' \mathscr{X} &= 0 \\ \mathscr{B} &= (\mathscr{X}' X)^{-1} \mathscr{X}' Y \end{aligned} \tag{9}$$

4.3 Iterative Algorithm

Subject clustering essentially consists of two subproblems. One is to determine K linear models and the other is to assign subjects to linear models appropriately. It is difficult to determine linear models and their memberships at the same time. However, it is not difficult to solve each of them separately. This observation motivates the iterative algorithm.

The basic idea of the iterative algorithm is to start from an initial solution of K linear models and then repeatedly perform the following two-step procedure:

- Given K linear models, assign I subjects to them appropriately (assignment step);
- Given a subject clustering solution, update K linear models (update step).

until a terminating condition is met.

Specifically, at the initialization step, we randomly pick K subjects and apply the OLS method to obtain K linear functions. If each section has no enough points to uniquely determine a linear regression function, we could group multiple subjects to generate one linear function or generate linear functions in a completely random fashion. In the assignment step, we assign each subject to the linear function which best fits the data. Suppose the linear model is $y = \alpha + \beta_1 x_1 + \ldots + \beta_J x_J$ and the data of subject i is $\{y_{it}, x_{ijt}\}$ with $t \in [1, T]$. The criterion of fitness is measured by

$$\sum_{t=1}^{t} (\alpha + \beta_1 x_{i1t} + \ldots + \beta_J x_{iJt} - y_i)^2.$$

In the update step, we apply the OLS method to each new cluster and obtain K new linear functions. We repeat the assignment and update steps till members of clusters are stable. The complete procedure is formally stated in Algorithm 2.

Algorithm 2 Iterative Algorithm for Subject Clustering

Input: $\{y_{it}, x_{ijt}, K\}$;
Output: $\{\alpha_k, \beta_k, P\}$;
1: Randomly pick K subjects and apply the OLS method to get K sets of $\{\alpha_k, \beta_k\}$.
2: Form k clusters by assigning each subject to the linear function which best fits its data and obtain their memberships P.
3: Apply the OLS method to each cluster and obtain K update parameter sets of $\{\alpha_k, \beta_k\}$.
4: If update parameter sets are same as previously, terminate, otherwise go to Step 2.

5 Experimental Study

In this section, we will conduct experiments on the statewide capital productivity data and the OECD gasoline demand data, which are provided in the Green's website,[1] to validate the effectiveness of our approach.

The general experimental setting is as follows. For a given panel data set with I subjects, J variables, and T periods, we divide the panel data set into two sets of periods. One set is employed as the training data to train the linear model and the other set is used to test the accuracy of the trained linear models. We compare our approach, referred to as *Selecting and Clustering* with the conventional pooling approach, referred to as *Pooling*, which pools all subjects together to infer a single linear model, the individualized approach, referred to as *Individual*, which considers I subjects belong to I different linear models and the conventional clusterwise linear regression approach, referred to as *Clustering*, which aims to cluster subjects into homogeneous groups with respect to all observed variables. For a fair comparison with the clusterwise linear regression approach, we also use the BIC information criteria to determine the number of clusters and use the same iterative algorithm to discover the underlying linear models. Essentially, the difference in our approach is that an additional variable selection phase, which is implemented through a simulated annealing search strategy, is added.

To evaluate prediction accuracy, we use the deviation as the evaluation criteria, which is defined as the following:

$$deviation(y, y') = \frac{|y - y'|}{|y|} \tag{10}$$

where y is the real value and y' is the predicted value. When $deviation(y, y') = 0$, y is accurately predicted. As the testing data contains more than one data record, the average value of deviations is used as the evaluation criteria to compare those four different prediction approaches.

[1] http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm

Table 1 Clustering result of productivity data

Cluster	States
1	CA, IL, IN, NJ, SC, TN, WI
2	AL, AR, DE, GA, IA, KY, ME, MA, MS, MO, NM, NH, NC, OR, PA, RI, SD, UT, VT
3	ID, MT, NV, ND, OK, TX
4	AZ, CO, CT, FL, KS, MD, MN, NM, WY,
5	LA, MI, NY, OH

5.1 Statewide Capital Productivity Data

The statewide capital productivity data set [2] consists of the data of the attributes of P_CAP (public capital), HWY (highway capital), $WATER$ (water utility capital), $UTIL$ (utility capital), PC (private capital), GSP (gross state product), EMP (employment), and $UNEMP$ (unemployment rate) for the lower 48 states in the U.S. from year 1970 to 1986. The research of interest is to predict the gross state product by examining other available attributes including public capital, highway capital, water utility capital, utility capital , private capital, employment, and unemployment rate. However, not all of the listed attributes might be useful for the prediction purpose. Another issue is that observations of 17 years are not enough to determine which attributes are significantly related to gross state product, the variable of interest. Our approach attempts to simultaneously identify attributes significantly related to the gross state product attribute and cluster 48 states into homogenous groups with respect to those identified attributes such that states in the same group belong to the same linear model.

For the producibility data set, we choose the first 13 years as the training data to train linear models and use the derived linear model to predict the next 4 years. As a result, our *Selecting and Clustering* approach selects three variables P_CAP (public capital), $UTIL$ (utility capital), and $UNEMP$ (unemployment rate) as explanatory variable for the response variable GSP (gross state product) and divide 48 states into five groups as show in Table 1. The prediction accuracy results are recorded in Table 2. Our *Selecting and Clustering* approach performs significantly better than the other three approaches. The performance of the *Clustering* approach is better than the *Pool* and *Individual* approaches, which demonstrates the importance of clustering subjects into homogenous groups. However, it considers all variables, including irrelevant variables, into linear models. The inclusion of those variables neither significantly explaining the dependent variable in pooled model nor contributing to the clustering identification could lead to bias when evaluating different clustering alternatives. Accordingly, the final clustering might not be efficient.

[2] http://pages.stern.nyu.edu/~wgreene/Econometrics/PanelDataSets.htm

Table 2 Comparison of prediction average deviation for productivity data

	Pooling	Individual	Clustering	Clustering and selection
Average Deviation	0.1139	0.0726	0.0061	0.0032

Table 3 Clustering result of gasoline data

Cluster	OECD countries
1	Belgium, Enmark, France, Greece, Italy, Netherland, Norway
2	Canada, Switzerland, Turkey
3	Spain
4	Sweden
5	Austria, Germany, Ireland, Japan, U.K.
6	U.S.A

5.2 Gasoline Demand Data

The world gasoline demand data[3] include 18 OECD (Organization for Economic Co-operation and Development) countries and have four attributes, namely, gasoline consumption per auto, real income per-capita, cars per-capita, and relative price of gasoline. We use the data with time span from 1960 to 1974 as the training data set and the data set from 1975 to 1978 as the testing data set. The gasoline price is the variable of interest.

As each subject contains only 13 observations, it is hardly to obtain precise estimates on the relationship between the gasoline price with other variables for each OECD country. This issue was identified by Baltagi [19]. He suggested to pool all countries together by assuming countries are fully homogenous. Vahid [7] made a further step and suggested that subjects could be clustered into groups such that subjects in the same group are homogeneous. In particular, he employed the agglomerative clustering method to iteratively group countries by evaluating the similarity between regression coefficients of subjects.

We make one more step further by suggesting to first select variables significantly correlated with the gasoline price and then use those selected variables to derive linear models. Our *Selecting and Clustering approach* discovers that the attribute of cars per-capita is significantly related to the gasoline price. According to the attribute of cars per-capita, 18 OECD countries are divided into six groups. Group memberships are recorded in Table 3. We derive a linear model for each group and use derived linear models to predict the next four years gasoline prices. The average prediction deviation of our approach and the results of the other three approaches are recorded in Table 4. The average prediction deviation of our approach is significantly less than other approaches. It strongly validates the effectiveness of our approach.

[3] http://pages.stern.nyu.edu/~wgreene/Econometrics/PanelDataSets.htm

Table 4 Comparison of prediction average deviation for gasoline data

	Pooling	Individual	Clustering	Clustering and selection
Average Deviation	0.9094	0.4379	0.4833	0.0462

6 Conclusion

In this chapter we showed the limitations of existing methods for panel data analysis and proposed a novel method to derive underlying linear models by identifying significantly related variables and grouping subjects into homogenous clusters. The BIC information criteria is employed to determine the number of clusters and the number of selected variables. It essentially balances the complexity of the model and the fitness of the model. This method is also applicable in research in Finance and Economics. Especially in a model specification with more variables for highly heterogenous subjects, previous approaches are very likely to have very large number of clusters which are practically less meaningful. In many real cases, the variable of interest is significantly correlated with only a few attributes. If one clusters subjects based on those few attributes, the true cluster memberships can be revealed. As a result, the number of clusters would be largely reduced and easier to interpret. To find the solution minimizing the BIC value, an efficient algorithm is coupled with a simulated annealing search strategy and an iterative algorithm. Expedients on real data sets showed that our approach performs significantly better than other existing approaches.

There are quite a few interesting extensions for this work. First, it might be interesting to extend our model to the unfixed clusterwise linear regression model, where different clusters are allowed to have different explanatory variables. Second, other than simulated annealing and iterative algorithms, there are many other well-known heuristic approaches. It is worth trying them to find more effective and more efficient algorithms to deal with our particular problem. Third, in this chapter, we used the subset selection to identify highly correlated variables. Other approaches such as the *LASSO* method are also able to discover parsimonious linear models. Exploring those issues will be our future work.

References

1. Baltagi, B.H.: Econometric Analysis of Panel Data, 3rd ed. Wiley, Chichester (2005)
2. Baltagi, B.H., Griffin, J.M.: Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline. J. Econometrics **77**(2), 303–327 (1997)
3. Durlauf, S.N., Johnson, P.A.: Multiple regimes and cross-country growth behaviour. J. Appl. Econometrics **10**(4), 365–384 (1995)
4. Kapetanios, G.: Cluster analysis of panel data sets using non-standard optimisation of information criteria. J. Econ. Dyn. Control **30**(8), 1389–1408 (2006)

5. Pesaran, M.H., Smith, R.: Estimating long-run relationships from dynamic heterogeneous panels. J. Econometrics **68**(1), 79–113 (1995)
6. Maddala, G.S., Wu, S.: Cross-country growth regressions: problems of heterogeneity, stability and interpretation. Appl. Econ. **32**(5), 635–642 (2000)
7. Vahid, F.: Clustering Regression Functions in a Panel. Monash University. Clayton (2000)
8. DeSarbo, W.S., Cron, W.L.: A maximum likelihood methodology for clusterwise linear regression. J. Classif. **5**(2), 249–282 (1988)
9. Baltagi, B.H., Griffin, J.M.: Gasolne Demand in the OECD: An application of pooling and testing procedures, gasolne demand in the OECD: an application of pooling and testing procedures. Testing for country heterogeneity in growth models using a finite mixture approach. J. Appl. Econometrics **23**(4), 487–514 (2008)
10. Castellacci, F.: Evolutionary and new growth theories. Are they converging? J. Econ. Surv. **21**(3), 585–627 (2007)
11. Castellacci, F., Archibugi, D.: The technology clubs: the distribution of knowledge across nations. Res. Policy **37**(10), 1659–1673 (2008)
12. Su, J.J.: Convergence clubs among 15 oecd countries. Appl. Econ. Lett. **10**(2), 113 (2003)
13. Zhang, B.: Regression Clustering, p. 451. IEEE Computer Society, Washington (2003)
14. Späth, H.: Algorithm 39: clusterwise linear regression. Computing **22**, 367–373 (1979)
15. Gaffney, S., Smyth, P.: Trajectory Clustering with Mixtures of Regression Models, pp. 63–72. ACM, New York, (1999)
16. Torgo, L., Da Costa, J.P.: Clustered partial linear regression. Mach. Learn. **50**(3), 303–319 (2003)
17. Ross, S.M.: Simulation, 3rd edn. (Statistical Modeling and Decision Science) (Hardcover). Academic Press, San Diego (2002)
18. Besag, J., Green, P., Higdon, D., Mengersen, K.: Bayesian computation and stochastic systems. Statist. Sci. **10**(1), 43–46 (1995)
19. Baltagi, B.H., Griffin, J.M.: Gasolne demand in the oecd: an application of pooling and testing procedures. Eur. Econ. Rev. **22**, 117–137 (1983)