Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operational Research

Estimating *R&D* spillover effects using panel data

Noortje Hopman (576523)

Supervisor:	Stan Koobs
Second assessor:	Dick van Dijk
Date final version:	1st July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This research examines the Pooled Lasso and Post Pooled Lasso estimation method proposed by Manresa (2016) for estimating interaction effects in the context of *R&D* spillovers between firms in the United States. The study utilizes an NBER matched Compustat-USPTO data set to obtain the results. Simulations are conducted to determine the circumstances under which the aforementioned methods yield the most accurate estimations and identify any limitations they may have. The paper addresses how the results can be interpreted, taking into account the limitations in the performance of the methods. Additionally, a Chow test is performed on a randomly selected subset of 10 firms to investigate whether the structural break in 1993 as suggested by Brown et al. (2009) is also observable within this specific data set. The Chow test demonstrates the presence of a structural break within the subset of firms, indicating a significant change in the magnitude of the interaction effects when comparing the period before 1993 with the period thereafter.

1 Introduction

Competition among firms has intensified as a result of a rapidly developing economy. To maintain a good competitive position, it is important for companies to have a unique strategy that sets them apart from others. If they fail to do so, they risk losing their market share as other companies take their place. Research and Development (*R&D*) has become a key method for enterprises to obtain the core competitiveness (Cao et al., 2022). Investing in *R&D* can lead to various benefits for a firm, such as the reduction of production costs, improvement of product quality, expansion of market share, and ultimately enhancing its overall market competitiveness. Due to for example the flow of human resources and information exchange between different firms, *R&D* spillovers arise. Spillover effects arise when for example firms use the know-how of another firm without the researching firm being able to control or influence the degree of this unintended knowledge transfer (Wölfl, 1998). Spillover effects are relevant not only among firms, but also in other contexts such as education, criminology, consumption, and productivity (De Giorgi & Pellizzari, 2014). One of the ways we observe spillover effects in education is when knowledge acquired in one subject contributes to improved understanding or performance in another subject. In addition to the unconscious spreading of knowledge, spillover effects also include the diffusion of strategies and different kinds of behavior. Research of Nilsson et al. (2017) for example has shown that environmental behavior is affected by the influence of spillover effects. In recent years, there has been increasing attention to the problem of competition and cooperation in *R&D* investments (Cao et al., 2022).

There is quite some literature on methods to estimate the magnitude of spillover effects between different firms. Most studies base their estimates on proxies for the structure of interaction and not on a data-driven estimate of the actual structure. Misleading results may arise as a consequence of this. In this research we treat the structure of interactions as unobserved and choose to estimate both the structure of interactions and the magnitude of the spillover effects. In order to do so we use a specific least absolute shrinkage and selection operator (Lasso) called the Pooled Lasso estimator to estimate the structure and the Post Pooled Lasso estimator to estimate the magnitude of the effects.

The main goal of this research is to find an answer to the following main research question:

‘How can the network of spillover effects between different firms be estimated?’ To get a clear understanding of how well the method performs in different settings, we begin by conducting several simulations. This approach aims to answer the first sub question: ‘In what circumstances does the method provide the most accurate estimates?’ The second sub question is: ‘What are the spillover effects among different firms in the United States between 1980 and 2001?’ To answer this question, we utilize an NBER marched Compustat-USPTO data set, containing information on sales and $R\&D$ investments of firms in the US during the period from 1980 to 2001. Lastly, the final sub question investigates whether the data set mentioned above indicates a structural break occurring in 1993, as suggested by Brown et al. (2009). In answering this sub question, we also examine the validity of the assumption made by Manresa (2016), which states that spillover effects are stable over time. To address this question, a Chow test is employed.

The simulations show that the sequential application of Pooled Lasso and Post Pooled Lasso generally leads to accurate estimates. In cases where the number of time periods is smaller than the number of firms, Pooled Lasso tends to provide better estimations compared to when Post Pooled Lasso is subsequently applied. When estimating the spillover effects between US firms using the described data set, we observe that spillover effects occur with varying magnitudes. Focusing on the first 8 companies in the sample, we find that high-tech firms in particular have large spillover effects on other firms. Additionally, the Chow test applied to a randomly selected sample of firms suggests the presence of a structural break in 1993. However, since the test was performed on a small portion of the sample, further research is necessary to determine if this conclusion holds true for the entire data set.

The remainder of this paper is structured in the following way. Section 2 contains a literature review providing more information on $R\&D$ expenditures, spillovers, and their changes over time. Additionally, some general information is provided about panel data and Lasso estimation. In Section 3, the data set used in this study is described, along with the implemented data cleaning procedures. Section 4 provides a detailed explanation of the models and methods employed to address the research questions. The results are presented in Section 5, showcasing the findings obtained from the analysis. Finally, Section 6 concludes the paper by summarizing the main outcomes and discussing some limitations of this research.

2 Theory

2.1 Literature review

Research and Development ($R\&D$) is a term to describe the process by which a company works to generate new knowledge that it might use to create new technology, products, services or systems. Companies in different sectors and industries conduct $R\&D$. A time series analysis done by Consult et al. (2008) investigates the relationship between GDP and $R\&D$. The research states that $R\&D$ intensities are temporarily influenced by the levels of GDP growth.

Lucking et al. (2019) study whether $R\&D$ spillovers have declined in the 21st century. They analyse panel data on US firms over the last three decades and find out that the magnitude of $R\&D$ spillovers remains as large in the second decade of the 21st century as it was in the mid 1980s. They do observe a temporary increase in positive $R\&D$ spillovers during the period of the digital technology boom in 1995-2004.

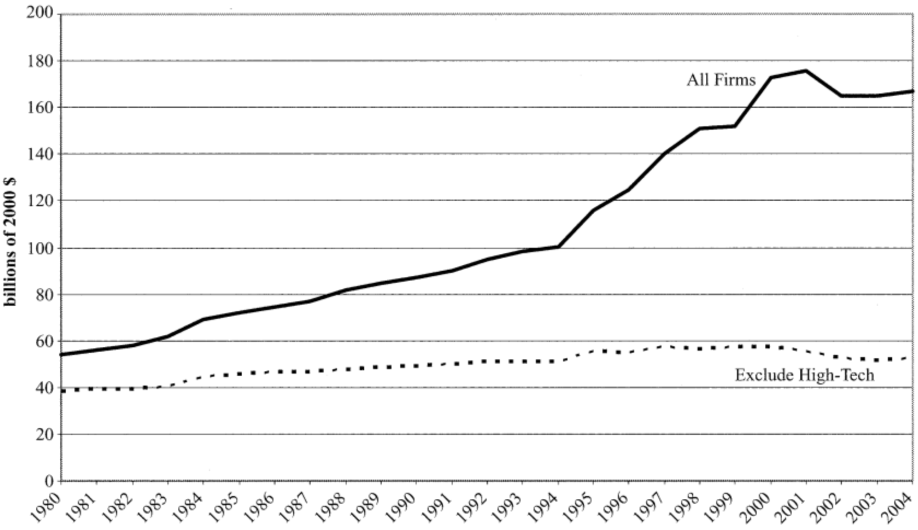

Figure 1: The solid line plots the sum of $R\&D$ for all publicly traded companies with coverage in Compustat (financial firms and utilities are excluded) over time. The dashed line plots the sum of $R\&D$ for firms in all industries except the seven high-tech industries with SIC codes 283, 357, 366, 367, 382, 384, and 737 (Brown et al., 2009).

The report of Brown et al. (2009) describes in detail how $R\&D$-expenditures developed between 1980 and 2004. Figure 1 plots the $R\&D$ investment in billions of 2000 dollars for all publicly traded firms listed in Compustat from 1980 to 2004. The dotted line is the level of $R\&D$ for all firms excluding seven high-tech firms. Splitting the group of firms in these two subgroups gives some interesting information on the share of high-tech firms on the development of $R\&D$ expenditures. We will describe shortly three things that can be concluded from looking at the results of Brown et al. (2009) in Figure 1. Firstly, economywide $R\&D$ starts accelerating in 1994 and ends around 2000. Secondly, the share of high-tech firms in $R\&D$ grew significantly in the period 1980-2004. In the last years of this time period, 2000-2004, approximately two-thirds of the total $R\&D$-expenditure was attributed to high-tech firms. Thirdly, the cycle in $R\&D$ between 1994 and 2004 is almost all due to the seven high-tech industries.

In her research, Manresa (2016) uses the same data as in this paper and assumes that the spillover effects remain stable over the period 1980-2001. The findings of Lucking et al. (2019) and Brown et al. (2009) collectively support the plausibility of a structural break occurring around 1993. Consequently, we conduct a Chow test in this paper to examine and test Manresa's assumption.

2.2 Panel data

Panel data, also known as longitudinal data, refers to a type of data that observes multiple variables over a specific period of time at regular intervals. It involves tracking the same group of individuals or companies throughout this time period. This type of data allows for the examination of both stability and changes over time. Time series data represents a one-dimensional case of panel data, where data is collected for a single variable. Another type of data is crosssectional data, which is collected at a single point in time. In this paper, we are dealing with panel data as we have data for a group of firms in the US, spanning the time period from 1980 to 2001.

2.3 Lasso estimation

As previously described in Section 1, we treat the structure of interactions as unobserved. Therefore, our starting point is to include interaction effects between all possible companies as potential components in our model. Nonetheless, including all the interaction effects in the model increases the likelihood of overfitting. An overfitted model contains an excessive number of variables given the data. Applying a standard regression method such as Ordinary Least Squares (OLS) to such a large set of explanatory variables (in this case all $N \times N$ possible interaction effects) could lead to an overestimation of how well the model performs. This phenomenon is called optimism bias (Ranstam & Cook, 2018). Lasso is a shrinkage and variable selection method that can be used to deal with the issue of overfitting in models by only selecting the most informative explanatory variables. Its objective is to strike a balance between accurately fitting the data (minimizing estimation errors) and preventing overfitting. This is achieved by adding a penalty term, λ, on the model parameters, which shrinks part of the regression coefficients towards zero. The Lasso estimator in its general form can be described as follows:

$$\widehat{\beta}_{Lasso} = \min_{\widetilde{\beta}} \left(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} \widetilde{\beta}_j)^2 + \lambda \sum_{j=1}^{p} \mid \widetilde{\beta}_j \mid \right). \tag{1}$$

In this formula n represents the number of individuals and p the number of parameters. The minimization process consists of two components. Firstly, there is the minimization of the sum of squared residuals, which represents the deviations between the estimated values generated by the regression model and the actual observations. Secondly, there is the inclusion of the absolute values of the regression coefficients, also referred to as the L1-norm. By minimizing the absolute values of the regression coefficients, the objective is to reduce their magnitudes as much as possible. One advantage of Lasso is that the L1-norm has the ability to selectively shrink certain regression coefficients precisely to zero. The degree of shrinkage depends on the penalty parameter λ. Variables with a coefficient of zero after shrinkage are excluded from the model. However, one of the limitations of Lasso is that it is sensitive to multicollinearity because it only randomly selects one variable of a set of highly correlated variables and ignores the rest of the highly correlated variables (Ogutu et al., 2012). Furthermore, the fact that Lasso shrinks the number of parameters also leads to another potential problem, namely shrinkage bias. This shrinkage bias will be further explained in Section 4 and is the reason why in this research a Post Pooled Lasso estimation is conducted after the Pooled Lasso estimation. Conducting simulations contributes to gaining insights into the circumstances in which this estimation method yields the most accurate estimates.

3 Data

We perform this research using an NBER matched Compustat-USPTO firm data set. This panel data set provides information on firm-level accounting data such as sales, employment and capital of firms in the United States over the period 1980-2001.

In our sample we include only those observations for which the values of the following variables are known: $code, year, SIC3, sales$ and xrd. The variable $SIC3$ refers to the Standard Industrial Classification (SIC) system. This is a coding scheme used to classify industries and businesses based on their economic activities. Access to variable $SIC3$ is desired as this offers information on the characteristics of various firms. Subsequently, once the results are obtained, an analysis can be conducted to examine how companies with specific characteristics relate to each other. The variable $code$ provides an identification code for each individual firm and is primarily used to differentiate between different companies, even if they may fall under the same SIC3. The variable $year$ represents the specific year associated with each data point. Given that we are conducting a panel data analysis, it is crucial to have knowledge of the corresponding time for all observations. In the regression analysis, the explanatory variable is represented by the variable xrd, which corresponds to the total expenditure in $R\&D$ in US-dollars. Conversely, the dependent variable is denoted by $sales$ and represents the total sales of a firm in US-dollars.

The full data set shows a different number of observations for each variable. Because $sales$ serves as dependent variable and has the fewest number of observations, we begin by removing all observations that do not have a value for $sales$. The observations we remove turn out to be the same observations that have missing values for the other variables. In other words, after excluding the observations without a value for $sales$, we have obtained a sample in which every observation contains values for all five variables mentioned above. Finally, we divide this sample into two sub samples to be able to conduct the Chow test: one containing all observations from 1980 till 1993, and another containing all observations from 1994 till 2001. Table 1 shows for every step in the process of data cleaning the corresponding number of observations per variable.

Step	**Description**	**Number of observations per variable**				
		code	year	sic3	sales	xrd
0.	Full data set	18209	18209	14084	13799	18209
1.	Remove all observations with missing values for sales and obtain sample 0	13799	13799	13799	13799	13799
2.	Include only observations from 1980-1993 and obtain (sub)sample 1	5145	5145	5145	5145	5145
3.	Include only observations from 1994-2001 and obtain (sub)sample 2	8654	8654	8654	8654	8654

Table 1: Steps to clean the data of missing values and make subsamples

Table 2 shows the number of observations over time. It can be seen that the number of observations remains relatively stable over time with an upward trend until 1993 and a subsequent downward trend after 1993.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Frequency	505	529	547	567	583	598	623	632	646	665	677	687

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	**Total**
Frequency	695	700	699	697	688	680	666	643	577	495	13799

Table 2: Distribution over time

Now, we will restructure the data set to create a coherent panel data set. The variable *code* serves as a means to group data points belonging to the same firm. We construct a matrix X that contains all values of xrd, structured so that each row corresponds to an individual firm and each column represents a different year. This results in an $N \times T$ matrix with N=725 and T=22 (1980-2001). Similarly, we create the matrix Y for the dependent variable $sales$. As Table 2 demonstrates, the number of observations varies across years because not every firm has data for each year from 1980 to 2001. Since one of our research questions aims to examine the constancy of spillover effects throughout the period of 1980-2001, we choose to include only those firms in our sample that possess values for all these years. Consequently, this leads to a sample of 348 firms.

Finally, there are also companies that have a value of zero for both xrd and $sales$ across all years. This suggests that either the data for these companies was inadequately recorded or the companies no longer exist or are inactive. To ensure the reliability of our findings, we decide to exclude those firms from our sample. As a result, the final sample comprises 274 firms. It should be noted that companies with one or more zero values across all 22 years will still be included in the sample as long as not all values equal zero. Table 3 summarizes the described steps in a structured way. The 274 firms in the final sample represent 94 different SIC3 codes. Traditionally $R\&D$-intensive industries, such as the Pharmaceutical Industry, the Electronic Industry and the Industrial Machinery industry are strongly represented in our sample. Some summary statistics on the sample can be found in Table 4.

Step	**Description**	**Number of different firms in sample**
1.	Restructure data set based on firm codes	725
2.	Remove firms with at least one missing observation in time period 1980-2001	384
3.	Remove firms where for the whole time period all xrd-values equal zero	274

Table 3: Steps to transform the data into a structured and complete panel data set

	Mean	**Standard Deviation**	**Minimum**	**Maximum**	**N**
x (*R&D*-expenditure in US-dollars)	162.4806	603.4154	0	8900	274
y (sales in US-dollars)	3977.94	12703.06	5	180557	274

Table 4: Summary statistics

4 Methodology

We use the following linear panel data model of spillover effects:

$$y_{it} = \alpha_i + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij} x_{jt} + \epsilon_{it}. \tag{2}$$

This equation shows that the sales of firm i at time t, y_{it}, can be affected by its own $R\&D$-expenditures at time t, x_{it}, but also by the $R\&D$-expenditures of other firms in the economy at time t, $x_{1t}, ..., x_{Nt}$. Note that $i = 1, ..., N$ and $t = 1, ..., T$. The estimate α_i is an firm-specific intercept, β_i is an firm-specific slope and γ_{ij} is a pair-specific parameters showing the effect of the characteristic of firm j on the outcome of firm i. All these γ_{ij}-values form together the matrix Γ with dimension $N \times N$, where the diagonal elements are β_i (also denoted as γ_{ii}). The interpretation of this γ_{ii} is the effect of firm i's own $R\&D$-expenditure on the sales of that certain firm i. The error term of firm i at time t is denoted by ϵ_{it}.

The micro-panel data we use consist of many more firms ($N = 274$) than periods of observation ($T = 22$). In this setting where $N >> T$, the data set becomes high-dimensional and this leads to certain consequences. The matrix of interaction effects to be estimated is of dimension $N \times N$, implying N^2 elements. Similarly, the explanatory variable X contains T different values for each firm, resulting in a dimension of $N \times T$. Given the large N in comparison to T, this situation leads to the problem of unidentifiability. To address the issue of unidentifiability, we assume sparse structures of interactions which leads to dimensionality reduction. In other words, we focus only on interaction structures where the number of connections for each individual is small. The number of sources of spillovers $\gamma_{ij} \neq 0$ is called s_i and has to be small relative to T in order to make our model perform well. This so-called sparsity assumption can be formally written as follows:

$$\sum_{j \neq i} \mathbb{1}\{\gamma_{ij} \neq 0\} = s_i << T \text{ for all } i. \tag{3}$$

By making this assumption, we aim to reduce the dimensionality of the problem to be able to identify relevant spillover effects. We use Lasso to find these relevant effects, which will be further discussed in Section 4.1.

In the remainder of this paper, we choose to consider x_{it} and y_{it} in deviations from their means over time $\bar{x}_i$ and $\bar{y}_i$. Consequently, the model presented in Equation (2) undergoes the following transformation:

$$(y_{it} - \bar{y}_i) = (\alpha_i - \alpha_i) + \beta_i(x_{it} - \bar{x}_i) + \sum_{j \neq i} \gamma_{ij}(x_{jt} - \bar{x}_j) + (\epsilon_{it} - \bar{\epsilon}_i). \tag{4}$$

$$\widetilde{y}_{it} = \beta_i \widetilde{x}_{it} + \sum_{j \neq i} \gamma_{ij} \widetilde{x}_{jt} + u_{it}. \tag{5}$$

Equations (4) and (5) illustrate the omission of the intercept term α_i as this effect is constant for a particular individual i, the representation of deviations from the mean using a tilde, and the construction of u_{it}.

The goal of our research is to estimate β_i and γ_{ij} consistently. To obtain these estimates we follow two steps, described in further detail below. Broadly speaking, we first perform a Pooled Lasso regression on the whole sample to estimate which interactions are non-zero and therefore are selected by the method. Our second step then is to perform a Post Pooled Lasso regression on the selected regressors to estimate the magnitude of the effects.

4.1 Pooled Lasso estimator

For every $i = 1, ..., N$ we perform the following Pooled Lasso estimation:

$$\widehat{\gamma}_i \in \underset{\gamma_{i1},...,\gamma_{iN}}{\arg\min} \frac{1}{T} \sum_{i=1}^{T} \left(\widetilde{y}_{it} - \sum_{j=1}^{N} \gamma_{ij} \widetilde{x}_{jt} \right)^2 + \frac{\lambda}{T} \sum_{j=1}^{N} \mid \gamma_{ij} \mid \phi_{ij}. \tag{6}$$

Equation (6) consists of two parts. The first part is the sum of squared errors and the second part is a penalization, involving a weighted sum of the absolute value of all spillover effects which is called the L1-norm. This method has two main advantages. Firstly, it produces sparse estimates by penalizing the absolute values of spillover effects. This leads to a relatively large number of spillover effects being set to zero. The fact that the method sets spillover effects to zero enables it to work effectively under the sparsity assumption of Equation (3). Secondly, using Pooled Lasso is computationally feasible due to the existence of efficient algorithms for solving the optimization problem described in Equation (6). The values of $\widetilde{y}_{it}$, $\widetilde{x}_{jt}$ and T in Equation 6 are given by the data. For λ and ϕ_{ij} we will perform some more calculations to find suitable values.

4.1.1 Determination of λ

In order to find a suitable value for λ we use the following formula, as suggested by Belloni et al. (2012):

$$\lambda = c \times (2 \times \sqrt{T} \Phi^{-1}(1 - v/2N)). \tag{7}$$

Φ denotes the standardized Gaussian cumulative distribution function. For parameter c we choose the constant 1.2 in line with the choice Manresa (2016) made in her paper. Parameter v is a pre-specified level of error, which we set equal to 0.05.

4.1.2 Determination of ϕ_{ij}

Determining a suitable value for ϕ_{ij} involves a series of steps. We use the iterative strategy proposed by Belloni et al. (2012), which works as follows. Initially, $\phi_{ij}^{2}{}^{(0)}$ is computed using the formula:

$$\phi_{ij}^{2}{}^{(0)} = \frac{1}{T} \sum_{t=1}^{T} \widetilde{x}_{jt}^2 \widetilde{y}_{it}^2 + \frac{1}{T} \sum_{t=2}^{T} \widetilde{x}_{jt} \widetilde{x}_{jt-1} \widetilde{y}_{it} \widetilde{y}_{it-1}. \tag{8}$$

The value of $\phi_{ij}^{(0)}$ is then obtained by taking the square root of $\phi_{ij}^{2}{}^{(0)}$. Subsequently, a Pooled Lasso estimation is performed individually for each i using Equation (6), with $\phi_{ij} = \phi_{ij}^{(0)}$.

After doing this for every $i = 1, ..., N$ separately, we have obtained all values of $\gamma_{ij}^{(0)}$. These $\gamma_{ij}^{(0)}$ values are subsequently employed to estimate $\widetilde{u}_{it}$ by employing the following expression:

$$\widehat{\widetilde{u}}_{it} = y_{it} - \sum_{j=1}^{N} \widehat{\gamma}_{ij}^{(0)} \widetilde{x}_{jt}. \tag{9}$$

The values of $\widehat{\widetilde{u}}_{it}$ obtained from this process serve as the initial point for calculating the subsequent ϕ_{ij} using the following formula:

$$\phi_{ij}^2 = \phi_j^2 = \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_{jt}^2\widetilde{u}_{it}^2 + \frac{1}{T}\sum_{t=2}^{T}\widetilde{x}_{jt}\widetilde{u}_{jt-1}\widehat{\widetilde{u}}_{it}\widehat{\widetilde{u}}_{it-1}). \tag{10}$$

Note that this is a natural estimator if ($\widetilde{u}_{i1}$ ... $\widetilde{u}_{iT}$) are independent and identically distributed (i.i.d.). While there is no certainty because $\widetilde{u}_{it}$ is unobserved, we make the assumption that they are i.i.d. and therefore choose to use the estimator of Equation (10).

After taking the square root of ϕ_{ij}^2, this ϕ_{ij} is used as input for the subsequent Pooled Lasso estimation. New values of $\widehat{\gamma}_{ij}$ are obtained, which are then utilized to calculate $\widehat{\widetilde{u}}_{it}$ according to Equation (9) once again. This iterative process continues until a new iteration no longer yields substantially different values for ϕ_{ij}. The final ϕ_{ij} obtained is utilized in the Pooled Lasso, where the corresponding γ_{ij} values ultimately determine the selection or exclusion of specific elements.

4.2 Post Pooled Lasso estimator

In this research, we use the Pooled Lasso estimator to identify the interactions between variables. However, there is a potential issue with this method called shrinkage bias.

Shrinkage bias can happen because of the way Pooled Lasso works. It shrinks some coefficients towards zero, which can lead to biased estimates for some variables. This bias occurs because Pooled Lasso has a preference for zero values, causing a downward bias. The strength of this bias depends on how much shrinking is applied, which is controlled by parameter λ. When the parameter is larger, more shrinking occurs, and the bias can be more pronounced.

To address this bias, we use an extra step called Post Pooled Lasso estimation. This technique helps correct the bias introduced by Pooled Lasso and improves the accuracy of the estimated values. Essentially, it refines the results obtained from Pooled Lasso to get a better understanding of the true relationships between variables. The Post Pooled Lasso estimator recovers the structure of spillover sources that were selected by the Pooled Lasso and simultaneously corrects for shrinkage bias. The Post Pooled Lasso estimator is defined as follows:

$$\widehat{\Gamma}^P = \underset{(\gamma_{i1},\ldots,\gamma_{iN}):\gamma_{ij}=0 \text{ if } j\notin\widehat{T}_i}{\arg\min} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widetilde{y}_{it} - \sum_{j=1}^{N}\gamma_{ij}\widetilde{x}_{jt})^2. \tag{11}$$

The set $\widehat{T}_i$ comprises all regressors selected by Pooled Lasso. All γ_{ij}'s that are not part of the set $\widehat{T}_i$ are set equal to zero beforehand. Solving the minimization of the sum of squared errors according to Equation (11) results in the estimator of matrix Γ which from now on we will call $\widehat{\Gamma}^P$ consisting of elements $\widehat{\gamma}_{ij}^P$.

4.3 Performance measures

After performing the calculations as described above, it is important to evaluate the performance of our model. With the simulated data, it is possible to compare the estimated matrix with the actual matrix. By comparing these findings, we can evaluate the model's performance and examine the effects of different input through various simulations. This provides an answer to one of our research questions, demonstrating how well the model performs and how different input influences its performance.

To test the accuracy of our estimation, we employ a few different performance measures. First of all, we use the commonly known Mean Squared Error (MSE), calculated as follows:

$$MSE = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \widehat{y}_{it}^{P})^2. \tag{12}$$

Note that $\widehat{y}_{it}^{P}$ represents the predicted value of y_{it} based on the estimated values of α_i and γ_{ij}. The error term ϵ_{it} has an expected value of zero since it is drawn from a standard normal distribution with a mean of zero and therefore disappears in the calculation of $\widehat{y}_{it}^{P}$. From Equation (2) it follows that the expected value of y_{it} can be calculated using the following formula:

$$\widehat{y}_{it}^{P} = \widehat{\alpha}_i + \widehat{\beta}_i x_{it} + \sum_{j \neq i} \widehat{\gamma}_{ij} x_{jt}. \tag{13}$$

The individual specific intercept α_i is estimated as follows:

$$\widehat{\alpha}_i = \bar{y}_i - \bar{x}_i^T \widehat{\gamma}_i^P. \tag{14}$$

The intuition behind this estimator is that it attributes the part of $\bar{y}_i$ that cannot be explained by $\bar{x}_i^T \widehat{\gamma}_i^P$ to the intercept α_i. The assumption of the error term ϵ_{it} being normally distributed with an expected value of zero is crucial in this context.

Another way to gain insight into the accuracy of an estimator is simply by examining the degree to which the estimates deviate from the true values. Since we want to compare matrices Γ with different dimensions in this paper, we choose to take the sum of the squared deviations and divide them by N. This way, we calculate the mean of the squared deviations, which we will abbreviate as MSD. In this paper, MSD_1 is used to represent the deviation of the Pooled Lasso Estimator $\widehat{\Gamma}$, while MSD_2 represents the deviation of the Post Pooled Lasso Estimator $\widehat{\Gamma}^P$. The formulas for these two statistics are as follows:

$$MSD_1 = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} (\gamma_{ij} - \widehat{\gamma}_{ij})^2. \tag{15}$$

$$MSD_2 = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} (\gamma_{ij} - \widehat{\gamma}_{ij}^{P})^2. \tag{16}$$

Finally, it is of interest to see whether the Pooled Lasso estimator correctly sets coefficients to zero. To gain insight into this, we utilize the following measure:

$$\text{Percentage correctly estimated zeros} = \frac{\#\widehat{\gamma}_{ij}^{P} \text{ estimated to zero correctly}}{\#\gamma_{ij} \text{ equal to zero}}. \tag{17}$$

When investigating the data set as described in Section 3, the true values of γ_{ij} are not known. Therefore, for evaluating those estimates, we rely solely on the MSE according to Equation (12). For the simulations on the other hand, the true values of γ_{ij} known and to analyse these estimates we also use the other performance measures.

4.4 Chow test

The Chow test is a statistical test that is used to test whether the coefficients estimated over one group of the data equal the coefficients estimated over another group. It is commonly used in time series analysis to test for the presence of a structural break, for example by Clark (2007) when testing a possible difference in market valuation of technology stocks before and after the crash. In our research, we use the Chow test to determine whether there are differences in interaction effects between companies during the period of 1980-1993 and 1994-2001. In order to do so we split up our sample in two subsamples. Let t run from 1 to 22, where $t = 1$ corresponds to the year 1980 and $t = 22$ to the year 2001. The first sub sample contains all data from period 1980-1993 and we denote the indices of this sample $t = \{1, ..., t_1\}$ where t_1 denotes 14, corresponding to the year 1993. The second subset contains all data from period 1994-2001 which indexes we denote as $t = \{t_1 + 1, ..., T\}$ where T denotes 22, corresponding to the year 2001. First, the Pooled Lasso estimator, as described in Equation 6, is applied to the entire sample. The γ_{ij}'s that are set to zero based on this estimation are also kept at zero in the subsequent regressions. We will execute the following regressions:

$$\widetilde{y}_{it} = \sum_{j=1}^{N} \gamma_{ij} \widetilde{x}_{jt} + u_{it} \text{ for all } t = \{1, ..., T\}. \tag{18}$$

$$\widetilde{y}_{it} = \sum_{j=1}^{N} \gamma_{1ij} \widetilde{x}_{jt} + u_{it} \text{ for all } t = \{1, ..., t_1\}. \tag{19}$$

$$\widetilde{y}_{it} = \sum_{j=1}^{N} \gamma_{2ij} \widetilde{x}_{jt} + u_{it} \text{ for all } t = \{t_1 + 1, ..., T\}. \tag{20}$$

If there is no structural break, all these regressions will lead to approximately the same estimates. To determine the presence of a structural break, it is therefore necessary to examine whether the estimates γ_{1ij} correspond to γ_{2ij}. In case of a structural break, these coefficients will significantly differ from each other. So, the Chow test actually tests the following hypotheses:

$$H_0 : \gamma_{1ij} = \gamma_{2ij} \text{ for all } i \text{ and } j.$$

$$H_1 : \text{otherwise.}$$

H_0 denotes the case of structural stability. If H_0 is rejected, we can therefore conclude there is no stability between the two selected time periods. To test whether the coefficients are jointly significant different from each other we use the following F-test in which N_1 denotes the number of observations of the first time period, N_2 the number of observations in the second time period and K the number of regressors.

$$F = \frac{(SSR_0 - SSR_1 - SSR_2)/K}{(SSR_1 + SSR_2)/N_1 + N_2 - 2K}. \tag{21}$$

The values for the Sum of Squared Residuals (SSR) can be computed using the following formula:

$$SSR = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \widehat{y}_{it}^{P})^2. \tag{22}$$

Note that SSR_0 denotes the sum of squared residuals of the whole data set (sample 0), SSR_1 the sum of squared residuals of the first group where $t = \{1 \ ... \ t_1\}$ (sample 1) and SSR_2 the sum of squared residuals of the second group where $i = \{t_1 + 1 \ ... \ T\}$ (sample 2).

The test statistic following from Equation (21) follows an $F(K, N_1 + N_2 - 2K)$ distribution. In this paper we choose to use a significance level of $\alpha = 0.05$. Learning from Pandey & Bright (2008) on the intuition behind degrees of freedom, it can be stated that degrees of freedom must always be positive. This leads to the following restriction:

$$N_1 + N_2 > 2K. \tag{23}$$

The number of regressors, K, is in the case of this research the number of selected parameters by the Pooled Lasso estimation and can be at most N^2. The total number of observations, $N_1 + N_2$ is $N \times T$. Knowing T to be much smaller than N a problem may arise. To avoid potential issues in the restriction on degrees of freedom, we choose to perform a Chow test on a small subset of all firms N. We choose to zoom in on 10 randomly selected firms, half of which belong to the high-tech industry and the other half do not. Section 5.3 will further elaborate on this choice.

5 Results

We use *R* version 4.0.5 to obtain the results described in this section. The codes used in Subsections 5.1, 5.2 and 5.3 are explained in Appendix A.1, A.2 and A.3 respectively.

5.1 Results simulated data

In order to evaluate the performance of the previously described model, we initially apply the model to a simulated data set. The data set is simulated in a simple way. First, we choose a value for N and T and simulate data for the variable x in the form of an $N \times T$ matrix. We fill this matrix with random numbers and vary the distribution from which we draw these numbers. Then, we generate a matrix Γ with dimensions $N \times N$, where we manually assign values at certain positions. A vector of length N is being constructed, which consists of individual-specific intercepts denoted as α_i. Finally, we simulate the data for variable y, again in the form of a matrix with dimensions $N \times T$, by multiplying Γ with X and adding for all i the individual-specific intercept α_i and the standard normally distributed error term ϵ_{it}. Per individual i we subtract the mean from the corresponding rows x_i and y_i to obtain $\widetilde{x}_i$ and $\widetilde{y}_i$, which we subsequently use in our model. Finally, the model's performance is assessed by comparing the matrix of actual γ-values to the values estimated by the method as described in Section 4.3.

We choose to describe a specific case in detail to illustrate the performance of our model. In this case, N equals 5 and T equals 15. Table 5 shows the manually constructed matrix

Γ. This Γ is constructed in such a way that every column contains approximately two zeros. The values for x are drawn from a continuous uniform distribution between 10 and 15, and the individual-specific intercepts α_i are drawn from a continuous uniform distribution between 3 and 5. The values of Y are calculated by multiplying the matrix Γ with X and for all i adding the individual-specific intercept α_i and the standard normally distributed error term ϵ_{it}.

$$\begin{pmatrix} 100 & 0 & 5 & 0 & 0 \\ 0 & 100 & 0 & 5 & 0 \\ 5 & 5 & 100 & 5 & 5 \\ 0 & 5 & 0 & 100 & 5 \\ 5 & 0 & 5 & 0 & 100 \end{pmatrix}$$

Table 5: Simulation example Γ-matrix

When this data has been simulated, the method described in Section 4 is applied step by step. This results in the Pooled Lasso estimator shown in Table 6. A value of zero in the matrix means that the parameter was not selected by the Pooled Lasso estimator. Then, Post Pooled Lasso is applied to the variables selected by the Pooled Lasso, resulting in matrix $\widehat{\Gamma}^P$ shown in Table 7.

$$\begin{pmatrix} 99.27 & 0 & 3.80 & 0 & 0 \\ 0 & 98.61 & 0 & 3.30 & 0 \\ 1.90 & 1.40 & 98.95 & 3.04 & 2.34 \\ 0 & 2.98 & 0 & 99.01 & 2.74 \\ 3.77 & 0 & 4.67 & 0 & 98.83 \end{pmatrix}$$

Table 6: Simulation example Pooled Lasso estimator $\widehat{\Gamma}$ [1]

$$\begin{pmatrix} 100.13 & 0 & 4.82 & 0 & 0 \\ 0 & 99.89 & 0 & 5.01 & 0 \\ 4.74 & 4.84 & 99.94 & 5.22 & 5.04 \\ 0 & 5.06 & 0 & 100.32 & 4.88 \\ 4.86 & 0 & 5.40 & 0 & 100.21 \end{pmatrix}$$

Table 7: Simulation example Post Pooled Lasso estimator $\widehat{\Gamma}^P$

It can be observed that the Pooled Lasso estimator selects the correct elements, resulting in a percentage of correctly estimated zeros of 100%. Additionally, as expected, the estimates of the Post Pooled Lasso are closer to the true values of Γ compared to the estimates of the Pooled Lasso. This is confirmed by the values of MSD_1 equaling 2.182 and MSD_2 equaling 0.022, calculated using Equations (15) and (16) respectively. We choose not to mention the values of the MSE because the estimator $\widehat{\alpha}_i$ turns out to not accurately estimate the values α_i. This can be seen in Table 8, showing the simulated values α_i versus the estimated values $\widehat{\alpha}_i$ for the simulation example discussed. Section 6 will further elaborate on this issue. The inaccurate estimates of $\widehat{\alpha}_i$ introduce noise in the MSE, because $\widehat{y}_{it}^P$ is calculated using $\widehat{\alpha}_i$ according to Equation 13. Therefore, we only report the values of the MSD, which provide a reliable indication of how accurately the estimator estimates the values of Γ.

[1] It took three iterations to obtain the weights ϕ_{ij} for the Pooled Lasso estimation

	i=1	**i=2**	**i=3**	**i=4**	**i=5**
Simulated value α_i	4.24	3.63	3.23	4.07	3.92
Estimated value $\widehat{\alpha}_i$	9.20	3.15	-8.59	3.41	-0.55

Table 8: Performance of estimator for α_i

We conduct several simulations to observe the model's performance in different cases. We systematically vary the number of individuals (N), the time period (T) and the number of γ_{ij}'s set to zero per individual. We report for all these simulations the values of MSD_1, MSD_2, and the percentage of correctly estimated zeros. Table 9 illustrates how these performance measures respond to changes in the data simulation approach. The values in that table are constructed by taking the average over two values obtained by running two times a simulation with corresponding N, T and number of γ_{ij}'s set to zero. To ensure fair comparisons between different cases, we change only one component at a time. The procedure to simulate values for x_{it}, α_i and ϵ_{it} remains consistent with the description provided above. Although the sizes and number of zeros in the Γ-matrices varies, the construction method remains the same, with diagonal elements set to 100 and off-diagonal elements set to zero or 50.

Case	**N**	**T**	**#γ_{ij}'s set to zero per firm**	MSD_1	MSD_2	**Percentage correctly estimated zeros**
1a	5	5	2	29.3007	26.8103	75%
1b	5	10	2	3.0426	0.0639	95%
1c	5	15	2	3.0971	0.7763	100%
1d	5	150	2	0.1108	0.0019	100%
1e	5	1500	2	0.0119	0.0003	100%
2a	50	5	20	176.5257	219.5798	93.6%
2b	50	10	20	29.7305	96.3712	86.9%
2c	50	15	20	21.7750	120.2689	91.9%
2d	50	150	20	0.3370	0.0025	97.5%
2e	50	1500	20	0.0208	0.0002	98.8%
3a	50	5	30	96.9685	103.8848	93.9%
3b	50	10	30	16.6473	49.5308	93.6%
3c	50	15	30	13.0413	14.7780	82.6%
3d	50	150	30	0.2223	0.0031	98.1%
3e	50	1500	30	0.0137	0.0001	100%

Table 9: Performance of different simulations

There are several insights to be gained from the results of Table 9. Firstly, there is a clear trend showing that the percentage of correctly estimated zeros increases as T (time period) increases. This trend holds true for cases 1, 2, and 3. A similar trend can be observed in the magnitudes of both MSD_1 and MSD_2: as T increases, these values decrease, indicating that the estimated values of Γ are closer to the true values. When comparing case 2 and case 3, it can be seen that a relatively higher number of zeros leads to lower values of both MSD_1 and MSD_2. In terms of the percentage of correctly estimated estimates, no clear trend is apparent. However, it can be observed that a relatively higher number of zeros at larger values of T (T=150 and T=1500) results in a higher percentage of correctly estimated zeros.

To assess the relationship between the Pooled Lasso estimates ($\widehat{\Gamma}$) and the Post Pooled Lasso estimates ($\widehat{\Gamma}^P$), it is interesting to compare MSD_1 with MSD_2. MSD_1 represents the performance of $\widehat{\Gamma}$, while MSD_2 represents the performance of $\widehat{\Gamma}^P$. In most cases, MSD_2 is lower than MSD_1, indicating that the additional step of Post Pooled Lasso estimation leads to improved estimates. The simulations where this is not the case have the characteristic that T is smaller than N. Based on these simulations, it can be concluded that the Post Pooled Lasso regression appears to be a meaningful additional step, particularly in cases where $T > N$.

5.2 Results empirical analysis

The same method used for the simulation is now applied to the data set described in Section 3. For the interpretation of the results, it is important to keep in mind the model of this research, namely Equation (2), repeated below:

$$y_{it} = \alpha_i + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij} x_{jt} + \epsilon_{it}. \tag{24}$$

In this equation, α_i, $\beta_i(=\gamma_{ii})$ and γ_{ij} have to be estimated. Pooled Lasso estimation and Post Pooled Lasso estimation are performed on the entire data set consisting of 274 firms. The value of the MSE, calculated using Equation (12) equals 9.15298×10^{14}. To maintain clarity, we choose to zoom in on the interaction effects between the first 8 companies to illustrate how the results can be interpreted. Table 10 shows the description of the industries to which these firms belong (Cognism, 2021).

SIC	**Description**
283	Medicinal Chemicals Botanical Products
367	Electronic Components Accessories
737	Services-Computer Programming
366	Telephone Telegraph Apparatus
357	Compputer Office Equipment
382	Laboratory Apparatus Furniture

Table 10: SIC-codes and description of industries of first 8 firms

Table 11 shows the estimated spillover effects $\widehat{\Gamma}^P$ between these 8 firms, and Table 12 the estimated firm-specific intercepts $\widehat{\alpha}_i$. The interpretation of this last table is the amount of sales in case of no expenditures in *R&D* for both the firm itself and all other firms. As mentioned earlier, the performance of this estimator is debatable and Section 6 will further elaborate on this. Note that although the remaining 266 rows and columns of Table 11 are not displayed, they are still included in the estimation process and contribute to the obtained estimates and MSE-value.

The first row of Table 11 represents a firm with SIC-code 283 and it can be seen that the $R\&D$-expenditures of a firm with SIC-code 366 influence the sales of that particular firm. One should be cautious about saying that a one-unit increase in the $R\&D$-expenditures of firm 366 results in a 16.80685 unit increase in sales for firm 283 due to the highly probable presence of omitted variable bias. This will be further discussed in Section 6. The firm with SIC code 367 is an example of a firm whose sales are influenced both positively and negatively by the $R\&D$ expenditures of other firms. Instead of looking at the firms that influence a particular firm, we can also examine the influence that a specific firm has on other firms. For instance, it can be observed that the $R\&D$-expenditures of the firm with SIC code 737 have a positive effect on the sales of its own firm and the firms with SIC codes 367, 366, and 357. When considering the descriptions of these firms in Table 10 this is not completely unexpected. The development in Services-Computer Programming has a significant influence because many companies utilize it (indirectly). This finding aligns with the statements of Lucking et al. (2019) and Brown et al. (2009), as described in Section 2.

	283	**367**	**737**	**366**	**366**	**357**	**382**	**382**
283	0	0	0	16.80685	0	0	0	0
367	4798.264	4.197076	19.79334	-1737.55	-7816.67	0	778.0774	-339.28
737	0	0	35.07492	-71.9285	0	0	0	0
366	0	0	4.328975	0	0	0	0	0
366	0	0	0	0	0	0	0	0
357	-1.19354	0	4.622841	0	0	0	0	0
382	0	0	0	0	0	0	10.20882	-0.18035
382	0	0	0	0	0	0	0	0

Table 11: First 8 rows of estimated $\widehat{\Gamma}^P$-matrix

SIC of firm	283	367	737	366	366	357	382	382
$\widehat{\alpha}_i$	-2847.28	24336660	129.40	-810.26	151.57	968.14	-2663.33	3791.10

Table 12: Estimated firm-specific intercepts $\widehat{\alpha}_i$

5.3 Chow test

As explained in Section 4, the possibilities to perform a Chow test are limited. To gain a good understanding of a possible literature-suggested structural break in 1993, we would need to perform a Chow test on the entire sample. However, due to Restriction 23 , this is not possible. Because literature suggests that the change after 1993 is mainly due to the contribution of the high-tech industries, we choose to zoom in on 10 randomly selected firms, half of which belong to the high-tech industry and the other half do not. To determine which firms fall under the high-tech industry, we use Brown et al. (2009) stating that firms with SIC codes 283, 357, 366, 367, 382, 384 and 737 belong to the high-tech industry. In this way, we investigate whether there is a structural change in the magnitude of the interaction effects when comparing the period before 1993 with the period after, on the scale of these 10 companies. Table 13 presents the firms selected using this approach. Alongside the SIC codes, industry descriptions from Cognism (2021) are included, and the final column indicates whether the firm belongs to the high-tech industry or not.

SIC	**Description**	**High-tech industry?**
211	Cigarettes	
267	Converted Paper & Paperboard Prods	
280	Chemicals And Allied Products	
331	Steel Works	
357	Computer & Office Equipment	Yes
366	Telephone & Telegraph Apparatus	Yes
367	Electronic Components & Accessories	Yes
384	Surgical & Medical Instruments & Apparatus	Yes
737	Service-Computer Programming	Yes

Table 13: SIC-codes and description of industries of selected firms

In this test $N_1 = N \times T_1 = 10 \times 14 = 140$, the total number of observations in time period 1980-1993 and $N_2 = N \times T_2 = 10 \times 8 = 80$, the total number of observations in time period 1994-2001. The value of K equals 76, the total number of regressors, in this case the number of non-zero values in the Pooled Lasso Estimator. Therefore the test statistic of this test will follow an $F(76, 68)$ distribution. The critical value is 0.678 on a significance level of α=0.05.

The calculated SSR-values, obtained using (22), are as follows:

- $SSR_0 = 169119825$
- $SSR_1 = 22610080$
- $SSR_2 = 806.71$

Plugging in these values in Equation (21) results in an F-statistic of 5.797514 which is greater than the critical value of 0.678. The p-value equals 2.395×10^{-12}, indicating strong evidence to reject the null-hypothesis of stability in magnitude of interaction effects between these 10 firms. In conclusion, it can be stated that the Chow test demonstrates a structural break in the magnitude of interaction effects for the 10 selected companies when comparing the period before 1993 with the period after.

6 Conclusion

This paper examines how a network of spillover effects between different firms can be estimated. The study utilizes the Pooled Lasso estimator and the Post Pooled Lasso estimator to obtain estimates of these effects. It investigates how these estimators work, their advantages and disadvantages, and the circumstances under which they lead to accurate estimates. Based on the simulation results, it can be concluded that the approach of sequentially applying Pooled Lasso and Post Pooled Lasso yields accurate estimates primarily when $T > N$. In cases where $T < N$, it is observed that the Pooled Lasso estimates generally outperform the Post Pooled Lasso estimates. This casts a critical perspective on the results obtained in Section 5.2, as the utilized data set has an N of 274 and a T of 22. The simulation has also clearly indicated that the employed estimator for α_i, as depicted in Equation (14), does not perform very well. Although this study did not have the opportunity to delve deeper into this issue, it would be a valuable subject of future research to further investigate the performance of this estimator $\widehat{\alpha}_i$. All of these factors together contribute to a possible explanation for the high MSE-value (9.15298×10^{14}) of the estimation of the whole sample.

The assumption made by Manresa (2016) in her paper, stating that the spillover effects are stable over time, is further investigated in this research. Due to the restrictions of the Chow test we chose to perform the test on a randomly selected sample consisting of five high-tech firms and five other firms. The Chow test shows with a p-value of 2.395×10^{-12} strong evidence of a structural break in 1993. This finding aligns with what the literature suggests, stating that the share of high-tech firms caused a structural break in $R\&D$-spillovers in 1993. However, it is crucial to interpret the conclusions of this Chow test with caution. It only indicates evidence of a structural break in the magnitude of the spillover effects. The analysis did not investigate whether there was a corresponding shift in the network's structure in 1993. Furthermore, the Chow test was limited to only ten companies, making it inappropriate to generalize the findings to the entire sample. Nonetheless, these findings show that within the subsample of 10 selected firms there was a structural break in magnitude of spillover effects and therefore form an indication to further investigate whether this structural break is also present in other settings. For example, it is interesting to conduct further research on potential changes in the structure of the network and the role of different industries within it. Additionally, exploring whether there are other potential breakpoints besides 1993 and investigating the underlying economic changes associated with them would be useful avenues of study.

This research has some other limitations that need to be addressed. Firstly, it was not possible to establish a comprehensive overview of the spillover effects among all firms included in the data set. This limitation arose from the data cleaning process, where a decision was made to remove all data associated with a company if any values were missing. Due to the lack of a proper procedure to handle missing values, this approach was adopted. To enhance future studies, it is recommended to employ interpolation techniques to estimate missing values. By implementing such techniques, it would be possible to retain a larger number of firms in the data set, thereby preserving valuable information for analysis purposes.

The simulation results were obtained by averaging data from two iterations. These iterations yielded approximately similar outcomes, justifying the decision to limit the analysis to two iterations. However, Driels & Shin (2004) suggests that conducting a Monte Carlo simulation with a significantly higher number of iterations could enhance the reliability and accuracy of the results. This is also plausible in our context because in our simulation numbers are randomly drawn from certain distributions. Therefore, there is a chance that the observed relationship between the size of N and T and the performance of the estimators is not a structural relationship but a result of coincidence. Although the anticipated conclusions of a Monte Carlo simulation are expected to align with the current findings, increasing the number of iterations in the simulation would therefore still be a valuable addition for further research.

Another important limitation of this paper pertains to omitted variables. It is evident that a firm's sales are influenced by numerous factors beyond its own *R&D* expenditures and those of its competitors. Therefore, we recommend further research that expands upon our methodology by incorporating a Double Pooled Lasso estimation, which includes other potential explanatory variables as part of the analysis. The steps involved in the Double Pooled Lasso estimation build upon the Pooled Lasso estimation and Post Pooled Lasso estimation conducted in this study. By expanding the model to incorporate multiple explanatory variables, the omitted variable bias is significantly reduced, enabling more direct interpretation of the estimated coefficients.

References

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.

Brown, J. R., Fazzari, S. M. & Petersen, B. C. (2009). Financing innovation and growth: Cash flow, external equity, and the 1990s r&d boom. *The Journal of Finance*, *64*(1), 151–185.

Cao, Y., Colucci, R. & Guerrini, L. (2022). On the stability analysis of a delayed two-stage cournot model with r&d spillovers. *Mathematics and Computers in Simulation*, *201*, 543–554.

Clark, E. (2007). Market valuation of technology stocks before and after the crash. *International journal of business*.

Cognism. (2021).
Retrieved from `https://www.cognism.com/sic-codes?utm_source=google_paid&utm_medium=paid_search&utm_campaign=emea_dsa&adgroupid=143765131599&utm_content=emea_dsa_allwebpages&utm_term=&device=c&network=g&gad=1&gclid=EAIaIQobChMI8LL037HI_wIVBPZ3Ch20mgWuEAAYASAAEgJVw_D_BwE`

Consult, I., Verbeek, A. & Lykogianni, E. (2008). A time series analysis of the development in national r&d intensities and national public expenditures on r&d. *Final Study Report for Specific Assignment*, *4*.

De Giorgi, G. & Pellizzari, M. (2014). Understanding social interactions: Evidence from the classroom. *The Economic Journal*, *124*(579), 917–953.

Driels, M. R. & Shin, Y. S. (2004). *Determining the number of iterations for monte carlo simulations of weapon effectiveness* (Tech. Rep.). NAVAL POSTGRADUATE SCHOOL MONTEREY CA DEPT OF MECHANICAL AND ASTRONAUTICAL ....

Lucking, B., Bloom, N. & Van Reenen, J. (2019). Have r&d spillovers declined in the 21st century? *Fiscal Studies*, *40*(4), 561–590.

Manresa, E. (2016). Estimating the structure of social interactions using panel data. *Unpublished Manuscript. CEMFI, Madrid.*

Nilsson, A., Bergquist, M. & Schultz, W. P. (2017). Spillover effects in environmental behaviors, across time and context: a review and research agenda. *Environmental Education Research*, *23*(4), 573–589.

Ogutu, J. O., Schulz-Streeck, T. & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *bmc proceedings* (Vol. 6, pp. 1–6).

Pandey, S. & Bright, C. L. (2008). What are degrees of freedom? *Social Work Research*, *32*(2), 119–128. Retrieved 2023-06-16, from `http://www.jstor.org/stable/42659677`

Ranstam, J. & Cook, J. (2018). Lasso regression. *Journal of British Surgery*, *105*(10), 1348–1348.

Wölfl, A. (1998). Spillover effects-an incentive to cooperate in r&d?

A Explanation programming code

In this section, a brief description is provided on how to use the code attached to this paper. The names of the documents with code in *R* correspond to the names of subsections A.1, A.2 and A.3. Please note that in addition to the short descriptions below, the code also includes detailed and extensive comments explaining each section of the code and what it accomplishes.

A.1 Code simulation

The first part of the code simulates the data. Note that there are some values assigned to N and T, but this values can of course be varied when doing different simulations. In line 34-139 we apply Pooled Lasso estimation to obtain the estimator for Γ, which we call coefsmatfinal. In line 141-148 we calculate the Post Pooled Lasso estimates and in the final lines 150-182 we calculate the different measures of performance.

A.2 Code empirical analysis

In line 5 the data is loaded. On the place of C:/Users/noort/Documents/Rotterdam/Erasmus/Bachelor 3/Blok5/Data/cleanedData.csv insert the working directory where the file with data called cleanedData.csv is saved to make the code work. The code consists of several parts. In the first part (line 1-70) we construct the data in a structural way and delete some observations as described in Section 3. In the second part (line 74-177) we apply Pooled Lasso estimation to find the estimator for Γ, which in the code is called coefsmatfinal. The third part (line 179-188) calculates the Post Pooled Lasso estimates and in the final lines 190-202 we execute some calculations to obtain the value of the MSE.

A.3 Code Chow test

The data can be loaded in the same way as described above. In the first part we randomly select 10 firms to perform the Chow test on and split these observations in two sub samples based on the time period. In the second part (line 74-177) we again apply Pooled Lasso estimation to find the estimator for Γ, which in the code is called coefsmatfinal. The third part applies the Post Pooled Lasso estimation method three times: for the whole sample, for observations in the first time period and for observations in the second time period. After doing this, line 231-268 take some steps needed to obtain the chow-statistic and p-value of the test.

To obtain the same selected firms as discussed in Section 5.3, use the following code instead of lines 67-77 in the attached code for Chow test:

```r
# Create empty matrices with dimensions 10xT
x_final <- matrix(NA, nrow = 10, ncol = T)
y_final <- matrix(NA, nrow = 10, ncol = T)
SIC_final <- matrix(NA, nrow = 10, ncol = T)

# Assign the selected rows to the matrices
x_final[1:5, ] <- x[c(213, 189, 51, 59, 233), 1:T]
y_final[1:5, ] <- y[c(213, 189, 51, 59, 233), 1:T]
SIC_final[1:5, ] <- SIC[c(213, 189, 51, 59, 233), 1:T]

x_final[6:10, ] <- x_hightech[c(45, 29, 59, 72, 9), 1:T]
y_final[6:10, ] <- y_hightech[c(45, 29, 59, 72, 9), 1:T]
SIC_final[6:10, ] <- SIC_filtered[c(45, 29, 59, 72, 9), 1:T]
```