Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Quantitative Finance

The Structure of R&D Spillovers across European Countries: A Simulation and Empirical Study

Lucas Crooijmans (569892)

Supervisor:	S.J. Koobs
Second assessor:	D.J.C. van Dijk
Date final version:	2nd 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 study contributes to the growing body of research on identifying spillovers with panel data. Previous literature has proposed methodology to identify individuals generating spillovers and their strength using panel data on outcomes and characteristics. This paper employs simulation to assess the performance of these methods, in addition to an empirical application thereof. This involves using the *Post Pooled Lasso* and the *Double Pooled Lasso* estimators, proposed by Manresa (2016), to evaluate European nationwide spillovers in the Research and Development sector. The data used is retrieved from Eurostat, the statistical office of the European Union. The performance of these estimators is first evaluated by simulation, with changes in sample size and with different weights in the Lasso regression. This part is to evaluate the robustness of these methods on differently sized data. In the empirical application, we find that R&D spillovers have a more complex structure across the European countries than is previously assumed. Furthermore, the *Double Pooled Lasso* estimator without using weights by iteration is optimal for low dimensional simulated panel data without apparent heteroskedasticity. Overall, this paper aims to provide a deeper understanding in previously proposed methodology on the identification of spillovers, and their applicability to a different context.

1 Introduction

Research and Development (R&D) activities play a crucial role in driving innovation and economic growth. In the process of conducting R&D, firms often generate knowledge and technological advancements that can spill over to other companies and contribute to their productivity and performance. These spillover effects have been widely recognized and studied in the field of economics. Understanding the nature and magnitude of R&D spillovers is essential for policymakers and firms seeking to enhance their innovation strategies and competitiveness.

In the seminal paper by Manresa (2016), she introduces an advanced methodology to estimate the structure and magnitude of R&D spillovers between American companies. By employing a regression analysis of total sales against R&D expenditures of other firms, Manresa (2016) investigates the influence of R&D activities of one company on the sales performance of others. This study takes a different direction and investigates international R&D spillovers in Europe.

While previous studies have examined R&D spillovers thoroughly already, they often rely on trade characteristics or non-trade related factors to estimate the effects. In recent studies, for instance those conducted by Moretti (2019) and Bianco (2012), foreign spillovers are aggregated, leaving the specific structure of spillovers unknown. In the context of international R&D spillovers, little research has focused on recovering the structure of interactions between countries. Therefore, there is a gap in knowledge regarding the structure of R&D spillovers between European countries and how it affects their economic outcomes. This gap motivates the central research question of this paper: '*How does the structure of R&D spillovers between European countries affect their economic outcomes?*'

In this paper, we focus on investigating the impact of different weights used in the *Post Pooled Lasso* and *Double Pooled Lasso* estimators proposed by Manresa (2016). These estimators provide methods to estimate social interactions (spillovers) using panel data. We perform these estimations on simulated data to enable comparison between the simulated and estimated values. For the simulation, we take the empirical data from Bloom, Schankerman and van Reenen (2013) as a starting point to generate simulated data that closely resembles the real-world application used in Manresa (2016). For the empirical application, we use data from Eurostat (2023) that covers a span of 22 years and includes 22 countries.

The methodology employed in this study involves the regression analysis of panel data to examine spillover effects. The output of each unit is regressed on its own characteristics as well as the characteristics of other units. We use the *Post Pooled Lasso* estimator, which involves a two-stage procedure of variable selection using Lasso regression followed by parameter estimation using ordinary least squares regression. Additionally, the *Double Pooled Lasso* estimator is presented for the case where auxiliary control variables are included. In this estimator, the Frisch-Waugh procedure is used to estimate the effects of control variables, after which the spillover effects are estimated through another *Post Pooled Lasso* regression. The empirical application in this paper focuses on examining the relationship between R&D expenditure and GDP using a modified version of the Cobb-Douglas production function. We estimate a model that regresses the logarithm of GDP on lagged R&D expenditure, as well as the spillover effects of lagged R&D expenditure from other countries. The model incorporates control variables for labor and capital stock, and we employ the *Double Pooled Lasso* estimator for estimation. To determine the optimal regularization parameter, we utilize Leave-One-Out Cross Validation.

Our analysis reveals insights into the structure of knowledge diffusion across countries and the impact of different weight choices on estimating R&D spillovers. The key results indicate that the *Double Pooled Lasso* estimator yields more accurate parameter estimates by considering control variables. Furthermore, our study opens avenues for further research into the structure of spillovers, as the results show that the R&D spillovers have a more complex structure than spillovers from geographical proximity or from technological distance.

This paper proceeds as follows. In Section 2, a overview of previous literature is given. The data used in this study is described in Section 3. In Section 4, the different modelling techniques are described in detail and the function of these models is explained. In Section 5, the results and possible implications of these results are discussed. We conclude and propose ideas for further research in Section 6.

2 Theory

2.1 Panel data regression

Panel data is a type of data that is collected by observing particular variables over a period of time at a regular frequency, capturing multiple observations per individual. Panel data is derived from a (usually small) number of observations over time on a (usually large) number of cross-sectional units like individuals, households, firms, or governments. The analysis of this datatype, panel regression, includes a range of statistical methods to analyze data that involves observations on the same individuals or units over multiple time periods. This analysis is also known as pooled time series analysis or longitudinal data analysis. There are various applications of panel regression in economics, including labor economics, health economics, development economics, and finance. In this paper, we look at international nationwide Research and Development (R&D) spillovers in the European market.

In econometric analysis, researchers often grapple with the challenge of selecting relevant variables while simultaneously controlling for model complexity and potential multicollinearity issues. Traditional linear regression models may yield suboptimal results in such scenarios Belloni, Chen, Chernozhukov and Hansen (2012), leading to biased estimates and limited predictive power. To address these concerns, a regularization technique called Lasso regression can be used to enhance the performance of a model.

2.2 Lasso regression

Lasso, short for Least Absolute Shrinkage and Selection Operator, was originally proposed by Tibshirani (1996) as an extension of the linear regression framework. It aims to simultaneously estimate the coefficients of the explanatory variables and perform variable selection by imposing a penalty on the absolute values of the coefficients. This penalty term shrinks certain coefficients to zero, effectively removing irrelevant variables from the model.

At the core of Lasso regression is the penalty term, which plays a vital role in both controlling model complexity and performing variable selection. In its most simple form, the penalty term is mathematically defined as the sum of the absolute values of the regression coefficients multiplied by a tuning parameter, denoted by λ. By introducing this penalty term, Lasso regression encourages sparse coefficient estimates by shrinking certain coefficients towards zero, effectively excluding irrelevant variables from the model. In essence, the penalty term acts as a trade-off between the goodness of fit and the complexity of the model.

The distinctive characteristic of Lasso regression lies in its ability to induce sparsity in the estimated coefficient vector, resulting in a more parsimonious model. This property is particularly valuable when dealing with high-dimensional datasets, where the number of potential predictors exceeds the number of observations. This makes it useful for analysis of panel data on spillovers, as the number of spillover parameters grows quadratically with the number of individuals, as seen in Manresa (2016). By encouraging sparsity, Lasso regression facilitates variable selection, providing us with a subset of important variables that have a substantial impact on the dependent variable.

One of the key advantages of Lasso regression is its capacity to handle multicollinearity issues, which arise when predictor variables are highly correlated with each other. In traditional regression models, multicollinearity can lead to unstable and unreliable coefficient estimates. However, the penalty term effectively encourages the selection of one variable over another, mitigating the adverse effects of multicollinearity, as shown in the simulation study by Altelbany (2021).

Moreover, the Lasso's penalty term can be tuned using cross-validation techniques, allowing us to strike a balance between model complexity and predictive accuracy. Chetverikov (2016) has demonstrated that in high-dimensional data, the use of cross-validation to select the appropriate penalty term is justified. In this paper, we use Lasso within a panel data framework. The panel data is estimated with the Pooled Lasso estimator (Manresa, 2016). Further details on the Pooled Lasso estimator are elaborated in Section 4.

2.3 R&D Spillovers

The research of Manresa (2016) includes an advanced method to retrieve estimates of R&D spillovers. Her research does not only estimate the structure of spillover effects that flows between companies, but also the amount of the spillovers that arise with the estimated network structure. Assuming that R&D expenditures enhance the total sales of firms, Manresa (2016) conducts a regression analysis of the total sales of a company against the R&D expenditure of all other companies in the sample. This approach enables the investigation of whether a company's sales are influenced by the R&D activities of other companies, and by what margins.

In this paper, we focus on the European market at national level. Other studies in this area include those conducted by Lumenga-Neso, Olarreaga and Schiff (2005) and León-Ledesma (2000). However, these studies assume the R&D spillovers either rely on trade characteristics like exports, or on other non-trade related characteristics. More recent studies on this subject, for instance by Moretti (2019) and Bianco (2012), also do not identify the structure of spillovers. These studies aggregate the foreign spillovers, leaving the structure of spillovers unknown. Within the research on international R&D spillovers, little to no studies have used methodology to recover the structure of interactions between countries, as is similarly done by Manresa (2016). From this gap in research, the following research question arises: '*How does the structure of R&D spillovers between European* *countries affect their economic outcomes?'*

Moreover, this research aims to investigate the effect of the penalty term and weights used in the *Post Pooled Lasso* and *Double Pooled Lasso* estimators on the estimation of R&D spillovers across European countries. The *Post Pooled Lasso* and *Double Pooled Lasso* estimators are methods proposed by Manresa (2016), which are used to estimate social interactions (spillovers) in panel data. In her study, she focuses on the application of this methodology to R&D spillovers between American firms. This study takes a different direction and estimates R&D spillovers on a nation-wide scale in Europe. Using nationwide data has a significant advantage over studying data on specific firms due to the availability of data. This is mainly because firms might not be as willing to disclose data due to the competitive market, as opposed to governments that often have freely available statistics.

By varying the penalty term and the weights used in the *Post Pooled Lasso* and *Double Pooled Lasso* estimators and examining their impact on simulated data of R&D spillovers, the study aims to determine what choice in penalty term and weights leads to the most accurate estimates. This provides insights into how both these estimators can be best applied in estimating R&D spillovers in the nationwide European market and may have implications for future research in this area. This leads to the following three sub-questions in the paragraphs below.

Sub Question 1: *What is the structure of R&D spillovers between European countries?* This question builds on the research from Lumenga-Neso et al. (2005), Moretti (2019) and Bianco (2012), by investigating the effect of international R&D spillovers between European countries. The previously mentioned studies focus on the total effect of all foreign R&D spillovers. This study zooms in on this subject and focuses on identifying the structure of spillovers between the European countries, to recognize patterns in the spillovers.

Within the framework of R&D spillovers, we employ two different models on panel data regression. The first model involves the estimation of spillover parameters exclusively, without the inclusion of explanatory control variables in the regression. Although this method is easier to perform, it can be subject to omitted variable bias (Belloni (2009), as potentially useful control variables are left out of the regression. As a second model, we include a set of control variables that affect all individuals in addition to the spillover variables. This is done to address the potential omitted variable bias that may arise in the first model specification. The specifications of both these models are elaborated on in Section 4.

Sub Question 2: *How does the choice of different weights in the Lasso regression impact the performance of Post Pooled Lasso and Double Pooled Lasso estimators?* The relevance of this question is to examine the effect when employing different weights on the penalty term of the Lasso estimator. By investigating whether distinct weights on the penalty term of the estimators improve the percentage of correctly selected variables, we aim to identify the optimal choice of weights that leads to the most accurate estimates. We test the different weight choices on simulated data, to enable comparison between the estimated values and the simulated values.

Sub Question 3: *How does the choice of the penalty factor in the Lasso regression affect the performance of the Double Pooled Lasso estimator?* In the empirical application of the study by Manresa (2016), the penalty factor is calculated with a fixed formula, where only the sample size of the data is used, without using the characteristics of the data used. This formula follows the methodology used by Belloni et al. (2012), where the formula is derived from statistical properties that hold under certain assumptions. However, the choice of the penalty factor in the Lasso regression can also be estimated on the data via cross validation, to adapt the penalty factor to the data used. Therefore, in the empirical application, we compare using cross-validation with using the fixed formula for lambda from Belloni et al. (2012) and Manresa (2016).

2.4 Cross-validation

Cross-validation is a powerful technique used to determine the optimal value for the penalty factor in Lasso Regression (Chetverikov, 2016). It helps strike a balance between model complexity and predictive accuracy by estimating how well the model generalizes to unseen data. The process of cross-validation involves splitting the available data into multiple subsets. One subset, also fold, is held out as a validation set, while the remaining folds are used to train the Lasso Regression model. The model is then evaluated on the validation set, using a performance metric such as the mean squared error, that we use in our cross-validation.

This procedure is repeated for each fold in a process known as k-fold cross-validation. For example, in 5-fold cross-validation, the data is divided into five equal parts, and the Lasso Regression model is trained and evaluated five times, with each fold serving as the validation set once. By repeating the cross-validation process for different values of the penalty factor, a range of mean squared errors is obtained. The penalty factor that yields the lowest average error across all folds is considered the optimal choice. Chetverikov (2016) found that using k-fold cross-validation is known to perform well in high-dimensional data, particularly when the amount of predictors is substantially larger than the number of time periods.

2.5 Cobb-Douglas framework

The Cobb-Douglas production function (Cobb & Douglas, 1928) is a widely used economic model that describes the relationship between inputs and output in the production process. The Cobb-Douglas production function has been employed to analyze various sectors of the economy, including national economies on a macroeconomic scale. It is a valuable tool in our analysis on the R&D spillovers, as it allows for the quantification of the impact of knowledge diffusion and technological advancements on economic output, as seen in Doi (2004). By incorporating R&D spillovers as an input, the production function enables researchers to assess the magnitude and significance of these spillovers on productivity and economic growth. This analysis provides insights for policymakers on the benefits of international R&D collaborations and knowledge-sharing initiatives, guiding the design of policies to foster innovation and enhance economic performance. On this nationwide scale, the Cobb-Douglas production function can be applied to measure the Gross Domestic Product (GDP) of a country. In Section 4.4, we provide a more detailed explanation of the Cobb-Douglas framework that is used.

3 Data

The data section is split up in two subsections. First, we discuss the origin of the data used in the simulation. Afterwards, the data used in the empirical application on European nation-wide R&D spillovers is presented in Section 3.2.

3.1 Simulation data

In the simulation, we use empirical data as a starting point to generate simulated data, to closely reflect the real-world application. The Research and Development case presented by Manresa (2016) employs data derived from Bloom et al. (2013). Descriptive statistics of the variables from Bloom et al. (2013) serve as the foundation for accurately generating the variables used in the simulation. Further details on the exact simulation of these variables can be found in Section 4.3.

3.2 Data on European R&D spillovers

For the empirical application on European nation-wide R&D spillovers, we retrieved data from Eurostat, the statistical office of the European Union. The database from Eurostat (2023) can be accessed through their website. From this database, data from 1995 until 2017 is retrieved to create a sample of 22 years. This database consists of 40 countries, most of which are in the European area.

However, upon examining the sample, we discovered that approximately 30% of the data points were missing. To improve the accuracy of the regression analysis, we decided to reduce the sample size to include only countries with more complete data. To achieve this, we established a criterion that no two consecutive data points could be missing. This effectively eliminated all countries with gaps of two or more years in their data from the sample. This left 22 European countries in the sample, with only 1.78% of all data missing. These data points, say d_t, are approximated by taking the average of the predeceasing value, d_{t-1}, and the succeeding value, d_{t+1}, also known as data imputation.

In total, four variables are extracted from the Eurostat database. The total amount of money spent on Research and Development by a country is measured using the Gross Domestic Expenditure on R&D (GERD) at the national level. The gross domestic product (GDP) of each country is used as measure for the total output in sales within the borders. The total employment of a country between the ages of 15 and 74 is taken as a measure of labor. Lastly, the physical capital of a country is measured with the consumption of fixed capital, as Blades and Meyer zu Schlochtern (1998) have shown that the consumption of fixed capital is the best candidate for international comparisons of total factor productivity. Some descriptive statistics can be seen in Table A1 in the Appendix.

4 Methodology

In Section 4.1 below, we describe the theoretical model to give a detailed overview on the regression of panel data. In Section 4.2, the estimation techniques are reported. As these techniques are mathematically complex, their origin and working are elaborated on extensively. In Section 4.3 and 4.4, the simulation and empirical application are discussed respectively.

4.1 Spillover Effects Model

In this section, the regression model of spillover effects with panel data is represented by the following equation:

$$y_{it} = \alpha_i + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij} x_{jt} + w'_{it}\theta + u_{it}, \tag{1}$$

where y_{it} is the output of unit i during time t. This regression includes an individual-specific intercept, denoted by α_i. The dependent variable is regressed on both its own characteristics, x_{it}, as well as that of other individuals, x_{jt}. So, evidently, the β_i is the effect that its own characteristics has on the output of that specific individual. The γ_{ij} is the spillover effect of unit j on unit i. For simplicity, β_i can also be seen as γ_{ii}. The w_{it} consists of the auxiliary explanatory variables, apart from the variables that generate the potential spillover effects between units. So, the effects of each auxiliary variable through θ are uniform across all individuals. The u_{it} represent the shocks of the regression that are uncorrelated with the explanatory variables. The spillover effects, γ_{ij}, are represented in the Γ-matrix.

4.2 Estimation techniques

This paper follows two different estimation procedures. First, we discuss the mathematics behind the *Post Pooled Lasso* estimator, where the model (1) is simplified to the case where $\theta = 0$. Afterwards, the *Double Pooled Lasso* estimator is explained in Section 4.2.2, in the case that $\theta \neq 0$.

4.2.1 Post Pooled Lasso

The *Post Pooled Lasso* estimator is contructed through a two-stage procedure. In the first step the significant variables are selected by a Pooled Lasso estimator, after which the parameters itself are estimated with Pooled OLS in step two. Manresa (2016) chose to use this *Post* estimation, as Belloni (2009) has shown that performing OLS after variable selection by Lasso regression can have the advantage of a smaller bias. On top of that, under certain assumptions, the *Post Pooled Lasso* estimator performs at least as well as using only Lasso, in terms of the rate of convergence. This gives the *Post Pooled Lasso* estimator a combination of the properties from Pooled Lasso and Pooled OLS. First, consider the minimisation function of the Pooled Lasso estimator:

$$\widehat{\Gamma} \in \underset{\Gamma}{\operatorname{argmin}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \sum_{j=1}^{N} \gamma_{ij} \tilde{x}_{jt} \right)^2 + \frac{\lambda}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \phi_{ij} \left| \gamma_{ij} \right|, \tag{2}$$

where $\widehat{\Gamma}$ represents the estimator of the spillover effects matrix. Here, the minimisation function is similar to that of OLS regression, as it minimises the sum of squared residuals. However, each parameter is penalised by the second part of (2), including penalty term λ. The ϕ_{ij} resembles the pair-specific weights that is investigated to affect the performance of the Lasso estimator. These weights are multiplied with the absolute value of the spillover effects, γ_{ij}, similar to the L1 norm in a normal Lasso regression (Tibshirani (1996)). Note that the dependent variable, $\tilde{y}_{it}$, and the explanatory variables, $\tilde{x}_{jt}$, are marked with a tilde. This implies that these variables are demeaned with respect to the mean of each individual, so $\tilde{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\tilde{x}_{it} = x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it}$ respectively. This is referred to as a within transformation, where the individual intercept is removed in a similar way to the fixed effects estimator.

The rows of the Pooled Lasso estimator in (2) represent the estimated coefficients for each individual. We compute the rows of this matrix independently by performing a Lasso regression on each unit individually. This effectively selects the right regressors, say $\widehat{T}_i$, for each individual. However, there still remains some shrinkage bias, as mentioned by Manresa (2016). To get rid of this bias, we perform Pooled OLS on the selected regressors by the Pooled Lasso estimation. See below the minimisation function of the Pooled OLS estimator with selected regressors, also *Post Pooled Lasso* estimator:

$$\widehat{\Gamma}^P = \underset{(\gamma_{i1},...,\gamma_{iN}):\gamma_{ij}=0, \text{ if } j\notin\widehat{T}_i}{\operatorname{argmin}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \sum_{j=1}^{N}\gamma_{ij}\tilde{x}_{jt}\right)^2, \tag{3}$$

where $j \notin \widehat{T}_i$ denotes that all regressors which are not chosen in the Pooled Lasso regression are set to zero. Consequently, the *Post Pooled Lasso* estimator only includes the regressors selected by the Pooled Lasso regression from the first step. Apart from that, this minimisation function is similar to plain Pooled OLS. The statistical properties of performing OLS after selecting regressors with Lasso are shown in Belloni (2009). Even if the Lasso fails to select exactly all relevant regressors, the performance of Post OLS regression beats single Lasso regression in terms of bias (Belloni, 2009). Although the Post OLS regression improves the shrinkage bias compared to using only Lasso regression, it does potentially result in an increased variance.

4.2.2 Double Pooled Lasso

In the case that $\theta \neq 0$, Manresa (2016) proposes the *Double Pooled Lasso* estimator. The following methodology includes auxiliary explanatory variables, also control variables, additional to the explanatory variables related to spillover. The first step of this method is similar to the technique used in Chamberlain (1992) to obtain orthogonal projections of both the output variable and the control variables on the spillover variables. This step is done by regression with the *Post Pooled Lasso* estimator. Afterwards, we use the Frisch-Waugh procedure to obtain an estimate for the control variables parameter. Once the effect of the control variables is estimated, we first subtract this effect from the dependent variable, to isolate the effect of the spillover variables. The last step is to regress the output (without the effect of the control variables) on the spillover variables by means of *Post Pooled Lasso* regression.

As stated above, the estimation of θ consists of two steps. First, both the dependent variable and the control variables are all separately regressed on $x_{1t}, ..., x_{Nt}$, similar to Chamberlain (1992). These regressions are done with the *Post Pooled Lasso* estimator. This captures the effects of all $\{x_{it}\}$'s on both the dependent variable, as well as on each of the control variables. There are D control variables in total. For each of these control variables, $w_{it}^1, ..., w_{it}^D$, we have the following regression:

$$w_{it}^d = \eta_i^d + \sum_{j=1}^{N}\lambda_{ij}^d x_{jt} + e_{it}^d, \tag{4}$$

where w_{it}^d represents the d'th control variable, η_i^d is the individual-specific intercept and the λ_{ij}^d stands for the effect of x_{it} on the d'th control variable. The e_{it}^d represents the shocks of the regression, uncorrelated with the explanatory variables. Similarly, the regression for y_{it} is:

$$y_{it} = \mu_i + \sum_{j=1}^{N} \nu_{ij} x_{jt} + v_{it}, \tag{5}$$

where y_{it} represents the output, μ_i is the individual-specific intercept and the ν_{ij} stands for the effect of x_{it} on the output, without the presence of control variables. The v_{it} represents the shocks of the regression, uncorrelated with the explanatory variables.

The residuals of (4) and (5) are used to obtain an estimate of the effect of the control variables, $\hat{\theta}$. Using the Frisch-Waugh-Lovell theorem (Lovell, 1963), we regress the demeaned residuals of (5), $\tilde{v}_{it}$, on the demeaned residuals of (4), $\tilde{e}_{it}^d$. This regression is done by Pooled OLS, where $\tilde{v}_{it}$ is calculated by $\tilde{y}_{it} - \widehat{\nu}_i \tilde{x}_t$, and $\tilde{e}_{it}^d$ is calculated by $\tilde{w}_{it} - \widehat{\lambda}_i \tilde{x}_t$.

Finally, the structure of interactions (including the respective magnitude of the spillover effects) is estimated by another *Post Pooled Lasso* regression. In this final regression, the dependent variable without the effect of the control variables ($\tilde{y}_{it} - \hat{\theta}\tilde{w}_{it}$) is regressed on all $\{x_{it}\}$'s. All *Post Pooled Lasso* regressions used in the *Double Pooled Lasso* estimator are calculated with the weights discussed in Section 4.2.3 below.

4.2.3 Choice of weights

A logical choice to allocate the weights, ϕ_{ij}, would be to use Heteroskedasticity and Autocorrelation Consistent (HAC) weights in Lasso regression on panel data. This would account for the presence of heteroskedasticity and autocorrelation in the data, as shown by the recent study of Babii, Ball, Ghysels and Striaukas (2021). The use of HAC weights in Lasso regression on panel data could be important as it helps to improve the efficiency of the estimator, by reducing the bias of the standard errors. In panel data, this can be useful as panel data often has presence of heteroskedasticity and autocorrelation. These characteristics can lead to biased standard errors. To mitigate this issue, it is recommended to allocate weights similar to the approach proposed by Babii et al. (2021). In this paper, we investigate if this approach is also feasible for small samples or short time frames. Therefore, we consider using these weights compared to no weight allocation at all, to show whether these weights can also be applied to small samples.

4.3 Simulation

This section outlines the methodology used to simulate data closely resembling the reality of American firms, based on the data used in Bloom et al. (2013). The simulation process involves four variables from the Cobb-Douglas production function: measure of output, labor, capital stock, and knowledge capital. Specifically, the output is measured in total sales of a company, the number of employees represents the labor variable, the net book value of property, plant, and equipment corresponds to the capital stock, and R&D expenditure represents the knowledge capital. The following steps are taken to simulate these four variables:

Initial Simulation Value: To begin the simulation, the minimum and maximum values from the descriptive statistics of the real data provided by Bloom are considered. The first value of each variable is randomly selected from a uniform distribution within the range of the minimum and maximum values. This step ensures that the simulated data starts within a similar range as the real data.

Time Period Variation: In each subsequent time period, variations are introduced to the simulated data to reflect the changing nature of the variables over time. For this purpose, a value from a normal distribution is added to each variable. The added value has a mean of 0 and a standard deviation equal to one-fifth of the starting value of the respective variable. This approach maintains the variables within a small range while allowing them to differ in each time period, thus emulating the dynamic nature of the real data.

Spillover Effects Matrix: Next, a spillover effects matrix is generated to capture the interactions between companies, mirroring the sparse structure outlined in the paper by Manresa (2016). To simulate this sparse structure, each company is assigned to one of 12 SIC codes at random, where companies receive spillover effects from all companies in their own industry (SIC code). In this way, each firm receives R&D spillover effects from only a limited amount of other firms. The effect of a company's own R&D is simulated using a normal distribution with a mean of 100 and a standard deviation of 25. On the other hand, the few spillover values originating from other companies are generated using a normal distribution with a mean of 10 and a standard deviation of 2.5. This distinction accounts for the fact that a company's own R&D expenditure typically has a greater effect on sales than those from other companies.

Sales Variable Generation: The sales variable is generated using three components: a firm-specific intercept, spillover effects multiplied by the R&D expenditures of the respective companies, and control variables multiplied by a randomly generated vector, θ. The values in θ are drawn from a uniform distribution ranging from 0 to 10. This approach incorporates various factors influencing sales and allows for random variations in the impact of the control variables.

By following this methodology, we aim to closely simulate data that resembles real data on American firms, based on the data provided by Bloom et al. (2013). The simulation process incorporates realistic variations in the variables over time, sparse spillover effects, and the influence of firm-specific and control variables on the sales variable.

4.4 Empirical Application

Consider the following formula for the nationwide Cobb-Douglas production function:

$$Y_{it} = A_{it} \cdot \left(C_{it}^{\theta_C} \cdot L_{it}^{\theta_L} \cdot K_{it}^{\beta_i} \cdot SK_{it} \right), \tag{6}$$

where the Y_{it} represents the GDP of a country. The subscripts i and t denote a specific country and time period, respectively. The capital stock is represented by C_{it}, which encompasses the physical assets and infrastructure available for production. The labor input is denoted as L_{it} and represents the quantity of the workforce in the country. Knowledge capital is captured by the variable K_{it}, which signifies the level of technological knowledge and expertise embedded in the production process. The knowledge spillovers, SK_{it}, reflect the effects of knowledge diffusion from all other countries combined. These variables are essential determinants of a country's economic growth and productivity, and their interactions are modeled using the Cobb-Douglas production function above.

The exponents in the Cobb-Douglas production function represent the output elasticities of the respective inputs, indicating the responsiveness of output (GDP) to changes in each input. The spillover effects, SK_{it}, are calculated as the product of all knowledge capitals of other countries:

$$SK_{it} = \prod_{j \neq i} K_{jt}^{\gamma_{ij}}, \tag{7}$$

where the γ_{ij} represents the knowledge spillover from country j to country i. So, the element γ_{ij} can be seen as the i'th country's output elasticity of the knowledge capital from country j. The logarithmic form of the Cobb-Douglas production function, can be expressed as follows:

$$\log(Y_{it}) = \log(A_{it}) + \theta_C \log(C_{it}) + \theta_L \log(L_{it}) + \beta_i \log(K_{it}) + \log(SK_{it}), \tag{8}$$

where each variable from (6) is put in logarithmic form. Taking the logarithm of the Cobb-Douglas production function allows us to transform the multiplicative relationship into an additive relationship, which simplifies interpretation and enables the econometric analysis discussed in Section 4.2.

For the empirical application in this paper we take the different elements from the previously mentioned model above, to the data mentioned in Section 3.2. Take the following formula, this time with variables applied to the R&D framework:

$$\log(GDP)_{it} = \alpha_i + \beta_i \log(GERD)_{i(t-1)} + \log\left(\sum_{j\neq i} \gamma_{ij} GERD_{j(t-1)}\right) + w'_{it}\theta + u_{it}, \qquad (9)$$

where GDP_{it} is the GDP of country i during time t. The logarithm of the dependent variable (GDP) is regressed on the lagged logarithm of both its own Gross Domestic Expenditure on R&D, $GERD_{i(t-1)}$, as well as that of other countries, $GERD_{j(t-1)}$. So, evidently, the β_i is the effect that its own GERD has on the GDP of the country. The γ_{ij} is the R&D spillover effect of country j on country i. For simplicity, β_i can also be seen as γ_{ii}. The w_{it} consists of the following two control variables: Labor and Capital stock. The u_{it} remain the shocks of the regression that are uncorrelated with the explanatory variables.

The GERD has been lagged by one time period, as the knowledge at time t can be proxied by the R&D expenditure at time $t-1$, similar to Manresa (2016). This is due to the dynamic relationship between R&D investment and knowledge that follows from the research done by Blundell, Griffith and van Reenen (1995). We estimate 9 by using the *Double Pooled Lasso* estimator.

We compare results from the fixed formula used in Belloni, Chernozhukov and Hansen (2014) and Manresa (2016) to using cross validation for the determination of lambda. The fixed formula is given by:

$$\lambda = c2 \cdot \sqrt{T}\Phi^{-1}\left(1 - \nu/\left(2N^2\right)\right) = 1.2 \cdot 2 \cdot \sqrt{22}\Phi^{-1}\left(0.05/\left(222^2\right)\right), \qquad (10)$$

where we pick c to be slightly above 1, and the pre-specified error ν is set to 0.05. The Φ represents the standardized Gaussian cumulative distribution. The derivation of this formula is cumbersome and the statistical background is explained in Belloni et al. (2012).

As the sample is rather small with only 22 time periods (see Section 3.2), using k-fold cross validation might not be optimal, according to Park (2010). The more suitable option that we use is Leave-One-Out Cross-Validation (LOOCV), as Park (2010) shows that it is more effective than k-fold, when dealing with limited data points. LOOCV operates by iteratively excluding one observation from the data, training the Lasso Regression model on the remaining data, and then evaluating the model's performance by predicting the omitted observation. This process is repeated for each observation in the data, resulting in a comprehensive assessment of the model's predictive accuracy. In our case, with a sample of only 22 observations, LOOCV can offer notable benefits as it maximises the training set size in each iteration. Therefore, it can reduce bias and yield a closer approximation of the model's generalization error.

4.4.1 Performance measures

For the empirical application, we analyse the effects of R&D on GDP, as well as the structure of interactions between countries. For the analysis on the effects of R&D on GDP, we calculate two performance measures: the elasticity of aggregate output with respect to the knowledge of country j ($\varepsilon^Y_{K_j}$), and the elasticity of aggregate output with respect to the knowledge of all countries (ε^Y_K). These performance measures allow us to quantify the impact of R&D on GDP at both the national and the aggregate level. The elasticity of aggregate output with respect to the knowledge of country j is calculated as:

$$\varepsilon^Y_{K_j} = \sum_{i=1}^{N} \gamma_{ij} \frac{Y_i}{Y}, \tag{11}$$

where $\varepsilon^Y_{K_j}$ is calculated by summing the product of the estimated spillover effects and the output of country i, divided by total output. This performance measure represents the change in aggregate output associated with a one-unit change in the knowledge of country j. The second performance measure on elasticity, ε^Y_K, is calculated as:

$$\varepsilon^Y_K = \sum_{i=1}^{N} \varepsilon^Y_{K_i}, \tag{12}$$

where all the individual elasticities are summed up to represent the elasticity of aggregate output with respect to the knowledge of all countries. This performance measure represents the change in aggregate output associated with a one-unit change in the knowledge of all countries.

For the analysis on the structure of interactions between countries, we use the estimated spillover effects from the *Double Pooled Lasso* estimator, denoted by $\widehat{\Gamma}$. This matrix contains the estimated coefficients for each independent variable in a LASSO regression model fit separately for each individual. The estimated coefficients represent the change in the dependent variable ($\log(GDP)_{it}$) associated with a one-unit change in the corresponding independent variable ($\log(GERD)_{i(t-1)}$), while holding all other independent variables constant.

5 Results

5.1 Simulation results

For the simulation, we focus on the comparison between using no weights, to using the weights discussed in Section 4.2.3. Tables 1 and 2 present the results of the *Post Pooled Lasso* and *Double Pooled Lasso* estimations on simulated data. Both tables compare the performance of two models: one that excludes weights and one that includes weights. The model performance is evaluated in terms of their ability to correctly estimate the zero and non-zero elements of the simulated spillover matrix, as well as the mean absolute difference between all estimated and simulated parameters.

Table 1: Results of the Post Pooled Lasso estimation on the simulated data.

	Excluding weights			**Including weights**		
	Zero β's	Non-zero β's	Difference	Zero β's	Non-zero β's	Difference
N = 5, T = 20	94.44%	100%	0.741	72.22%	85.71%	1.703
N = 5, T = 100	94.44%	100%	0.498	88.89%	100%	0.598
N = 5, T = 1000	100%	100%	0.307	100%	100%	0.307
N = 20, T = 20	97.09%	78.57%	0.793	63.08%	66.07%	2.247
N = 20, T = 100	100%	100%	0.293	92.15%	98.21%	0.375
N = 20, T = 1000	100%	100%	0.268	100%	100%	0.268
N = 1000, T = 20	98.49%	2.97%	1.358	98.13%	2.19%	10.12

Note: The zero and non-zero β columns represent the percentage of correctly estimated β's of the zero and non-zero elements of the simulated spillover matrix. The difference column is the mean absolute difference between all estimated and simulated parameters.

When comparing the results from the *Post-* and *Double Pooled Lasso* estimations, we can see that for all combinations of N and T, both models perform similarly. Their ability to select the right regressors is almost equal in strength, as one can note that both these models have similar percentages of correctly estimated zero and non-zero elements in $\widehat{\Gamma}$. However, there is one main difference in performance between the *Post-* and *Double Pooled Lasso* models. The mean absolute difference is in all cases lower for the *Double Pooled Lasso* model. This is according to expectation, as the output variable is dependent on the control variables by nature of simulation. This dependency on the control variables gives the *Post Pooled Lasso* an omitted variable bias, as the control variables are not accounted for and leave a bias to the parameters on spillovers, as discussed in Section 2.3. On the other hand, the *Double Pooled Lasso* estimator does account for the dependency on control variables, giving it more accurate parameters for the spillover matrix.

The results of the simulation analysis reveal an interesting finding regarding the performance of the model that includes weights compared to the model without weights. Surprisingly, the Lasso model without weights outperforms the model with weights in terms of selecting the right regressors and on estimation accuracy.

One possible explanation for this outcome is that the simulated data used in the analysis does not exhibit autocorrelation or heteroskedasticity. The HAC weights are

Table 2: Results of the Double Pooled Lasso estimation on the simulated data.

	Excluding weights			**Including weights**		
	Zero β's	Non-zero β's	Difference	Zero β's	Non-zero β's	Difference
N = 5, T = 20	94.44%	100%	0.681	77.78%	85.71%	1.467
N = 5, T = 100	94.44%	100%	0.465	94.44%	100%	0.392
N = 5, T = 1000	100%	100%	0.300	100%	100%	0.300
N = 20, T = 20	97.38%	76.79%	0.761	65.41%	64.29%	2.150
N = 20, T = 100	100%	100%	0.284	91.57%	96.43%	0.407
N = 20, T = 1000	100%	100%	0.265	100%	100%	0.265
N = 1000, T = 20	98.49%	2.97%	1.358	98.13%	2.23%	8.979

Note: The zero and non-zero β columns represent the percentage of correctly estimated β's of the zero and non-zero elements of the simulated spillover matrix. The difference column is the mean absolute difference between all estimated and simulated parameters.

specifically designed to address these issues by adjusting the standard errors of the estimated coefficients to account for potential serial correlation and heteroskedasticity in the data. However, in the absence of these characteristics, the application of these weights may introduce unnecessary noise and bias into the estimation process.

Furthermore, the Lasso model is inherently designed to handle high-dimensional data and perform variable selection by shrinking the coefficients of irrelevant variables to zero. Introducing these weights may introduce additional complexity and potential distortions to the variable selection process, leading to less accurate coefficient estimates and a reduction in the model's predictive performance. It is worth noting that the superiority of the Lasso model without weights over the model with HAC weights in this specific context does not diminish the importance of accounting for autocorrelation and heteroskedasticity in real-world applications. In empirical studies with actual data, these issues are prevalent and can significantly impact the reliability of coefficient estimates. In such cases, the inclusion of weights is crucial for obtaining valid inference and accurate estimation results.

As each of the simulated companies gets allocated to one of 12 industries (see Section 4.3), the models with a low N, are relatively more sparse than a model with a high N. The results of the analysis indicate that when N is low, such as $N = 5$ or 20, the Lasso model predicts the sparse spillovers relatively well. However, when N increases to 2000, representing a relatively dense model, the model performs poorly in predicting the non-zero spillover parameters. This outcome can be attributed to the density of spillover effects when N is high. In the case of a low number of companies per industry, the spillover effects among individuals are expected to be sparse, meaning that only a small subset of individuals have significant spillover impacts on others. In such scenarios, the Lasso model's variable selection property effectively identifies and captures the relevant spillover relationships, leading to accurate prediction of the non-zero spillover parameters. However, with 2000 companies, the number of companies per industry increases substantially, where the model faces a significant increase in relative amount of spillovers per firm. With a larger number of significant variables to consider, the Lasso model encounters a higher noise-to-signal ratio, making it challenging to accurately estimate all the non-zero spillover parameters. This issue arises when the spillover effects are dense, or in this case when substantial number of individuals have significant spillover impacts on others. Another factor contributing to the poor prediction of non-zero spillover parameters in the dense model is the potential limitation in the sample size.

5.2 Results of international R&D spillovers in Europe

For the estimation of the structure of interactions between the European countries, we present two tables. Table A4 in the Appendix uses the fixed formula for the penalty term, while Table 3 uses LOOCV to determine the penalty term. Tables 3 and A4 show the estimated spillover matrix on the international R&D spillovers between European countries. The values in the table represent the effect of R&D in one country on the GDP of another country. A positive value indicates that an increase in R&D in one country has a positive effect on the GDP of another country, while a negative value indicates a negative effect.

In Table A4 in the Appendix we see that there are estimated to be only five spillover effects across countries. In Manresa (2016) it is shown that the fixed formula has useful statistical properties for when T is large. However, combining this specific penalty term with the HAC weights can lead to an underselection of spillover effects if T is small, for further explanation on this see Manresa (2016). As the results show only five spillover effects between all 22 countries, using the fixed formula for the penalty term might not be useful for analysis on the structure of interactions of the R&D between these European countries. Therefore, we use cross validation for the lambda, LOOCV in specific, to look at the more noticeable spillover effects in Table 3.

Based on the data provided in Table 3, it appears that there are complex interactions between the R&D of these countries. Some countries have strong positive spillover effects on other countries, while others have negative effects. For example, an increase in R&D in Austria (AT) has a positive effect on the GDP of Bulgaria (BG), Czech Republic (CZ), and Poland (PL), while it has a negative effect on Italy (IT). Similarly, an increase in R&D in Belgium (BE) has a positive effect on Hungary (HU), Poland (PL), Portugal (PT), Sweden (SE), and Slovenia (SI), while it has a negative effect on Greece (EL).

Table 3: The estimated spillover matrix on the international R&D spillovers using LOOCV for the penalty term.

	AT	**BE**	**BG**	**CZ**	**DE**	**DK**	**EL**	**ES**	**FI**	**FR**	**HU**	**IE**	**IT**	**LT**	**NL**	**PL**	**PT**	**RO**	**SE**	**SI**	**SK**	**UK**
AT	0.48	0	1.5	1.3	0.2	0.37	0	0	0	0.19	0	0	-0.035	0	0	0.71	0	0	0	0.19	0	0
BE	0.0075	0.42	0	-0.23	0	0	-0.87	0	0	0	0.3	0	0	0	0	1.7	0.82	0	0.67	0.49	0	0
BG	0	0	0.086	0	0	0	0	0	0	0	0	0	0	0	0	-0.27	0.16	0	0	0	0	0
CZ	0	0	0	-0.29	0	0	0	0	0	0	0	0	0	0.35	0	0	-0.2	0	0	0	0.43	0
DE	0	0	-1.4	-1.3	0.45	0	0	-0.41	0	0	0	0	0	0	-0.18	0	-0.96	-1.7	0	0	0	-0.93
DK	0	0	0	0.2	0	-0.099	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.24
EL	0	0	0	0	0	0	0.18	0.16	0	0	0	-0.018	0	0	0.21	0	0	0	-0.45	-0.026	0	0
ES	0	0.3	0	0	0	0	0.52	0.53	0	0.16	0.41	0.052	0.2	0	0.1	0	0.4	0	0.41	0.51	0	0
FI	0	0	0.3	0	0	0.12	0	0	0.23	0	0	0	0	0.15	0.07	0.49	0.22	0.56	0	-0.12	0.62	0.17
FR	0	0	0	1.1	-0.4	0	0	0	0	-0.13	0	0	0	0	0	-1	0	0	-0.54	-0.58	0.68	-0.2
HU	0	0	0	0	0	0	0	0	0	0	0.29	0	0	0	0	0	0	0	0	0	0	0
IE	0	0	0	0	0	0	0	0	0	0	0	0.43	0	0	0	0	0	0	0	0	0	0
IT	0	0	0	0	0	0	0	0	0	0	0	0	0.072	0	0	0	0	0	0	0	0	0
LT	0	0	0	0.14	0	0	0	-0.049	0	0	0	0	0	0.14	0	0	0	0	0	0.078	0	0
NL	0	0.064	0	0	0.32	0.16	0	0.74	0.34	0.24	0.59	1.1	0	0.54	0.4	0	0.25	2.5	0.44	0	0.82	1.4
PL	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.26	0	-0.14	0	0	0	0	0	0
PT	0	-0.13	0	0	0	0	0	0	0.016	0	0	0	0	0.08	0	0	-0.033	0.13	0	0	0	0.09
RO	0	0	0	0	0	0	0	0	0.043	-0.014	0	0	0	0.072	0	0.15	0	0.47	0	0	0.28	-0.12
SE	0	0	0	0	-0.19	0	0.17	0	-0.091	-0.084	0	0	0	-0.11	0	-1.3	0.12	-1.1	-0.16	0	-0.86	0
SI	0	0	0	0	-0.01	0	0	-0.21	0	0	-0.41	-0.53	0	0	0	0	-0.3	0	0	-0.21	-0.37	0
SK	0	0.0058	0.1	0	0	0	0	0	0	0	-0.12	0	0	0	0	0	0	0	0	0	-0.04	0.065
UK	0	0	0	0	0	0	0.28	0	0.22	0.088	0.038	0.57	0.23	0.39	0	0.43	0	0.46	0.17	0	-0.12	0.33

Note: The countries are abbriviated by two-letter country codes defined by ISO 3166-1. See Table A2 in the Appendix for their respective country names.

These results suggest that there are complex relationships between the R&D of these countries and their GDPs, which can not be easily explained by geographical proximity or by technological distance of these countries, as done in León-Ledesma (2000) and Moretti (2019).

Table A3 shows the elasticities of aggregate output with respect to the knowledge of country j, $\varepsilon^{Y}_{K_j}$. These values can be interpreted as follows: a positive value for $\varepsilon^{Y}_{K_j}$ indicates that an increase in the knowledge of country j has a positive effect on aggregate output, while a negative value indicates a negative effect.

Based on these results, we see that some countries have strong positive effects on aggregate output, while others have negative effects. For example, an increase in the knowledge of Austria (AT) has a positive effect on aggregate output, while an increase in the knowledge of Belgium (BE) has a negative effect. Similarly, an increase in the knowledge of all countries has a positive effect on the output of Austria (AT), while it has a negative effect on the output of Belgium (BE). The elasticity of aggregate output with respect to the knowledge of all countries, ε^{Y}_{K}, is estimated at approximately 0.01328. This indicates that a 10% increase in the knowledge of all countries gives a 0.13% increase in aggregate GDP.

6 Conclusion

This paper addresses the research gap in the study of R&D spillovers at the national level in the European market. Our analysis is built upon previous studies that focus on the total effect of foreign R&D spillovers, without distinguishing the specific structure of interactions. By employing panel data regression analysis, we estimate the spillover parameters while also controlling for additional factors that affect all individuals.

The central research question of this study was to understand the structure of R&D spillovers between European countries and determine the optimal choices for the penalty term and the weights in estimating these spillovers. We addressed this question by employing panel data regression models and conducting simulations to compare the performance of different models and estimators.

The main findings of this study indicate that both the *Post Pooled Lasso* and *Double Pooled Lasso* models perform similarly in terms of correctly selecting relevant variables. However, the *Double Pooled Lasso* estimator, which incorporates control variables, yields more accurate parameter estimates for the spillover matrix. This suggests that accounting for the dependency on control variables is crucial to reduce bias in estimating R&D spillovers. These findings emphasizes the importance of tailoring the modeling approach to the specific characteristics of the data and highlights the need to consider the appropriateness of including additional weights and adjustments based on the nature of the data and the research objectives.

In response to our research question, we can conclude that the structure of R&D spillovers between European countries exhibits certain patterns and dynamics that do not rely on trade or geographical proximity, which needs further research to identify what does drive these spillovers. An interesting topic for further research in this area would be to investigate what drives these complex patterns. Knowledge diffusion within the region is influenced by various factors, and considering control variables in estimating spillovers provides more accurate results.

Practically, these findings have implications for policymakers by identifying patterns and mechanisms of knowledge diffusion, so policymakers can make informed decisions to maximize the positive impacts of R&D spillovers on national economies. Theoretical implications of this study lie in the advancement of knowledge regarding the estimation of R&D spillovers and their structure. These findings can further improve existing theoretical frameworks related to knowledge diffusion.

Future studies could expand the analysis to include additional countries or regions beyond Europe to gain a broader perspective on R&D spillovers. Furthermore, it would also be interesting to investigate how these interactions have changed over time and how they are influenced by factors such as trade policies, economic integration, and technological developments. Additionally, an interesting topic is to validify the usage of HAC weights where there is heavy heteroskedasticity or autocorrelation present in simulation.

References

Altelbany, S. I. (2021). Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: A simulation study. doi: 10.34260/JAEBS.517

Babii, A., Ball, R. T., Ghysels, E. & Striaukas, J. (2021). *Machine learning panel data regressions with heavy-tailed dependent data: Theory and application.*

Belloni, A. (2009). Least squares after model selection in high-dimensional sparse models. *Econometrics: Econometric & Statistical Methods - General eJournal*. doi: 10.3150/11-BEJ410

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. doi: https://doi.org/10.3982/ECTA9626

Belloni, A., Chernozhukov, V. & Hansen, C. (2014, May). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29-50. doi: 10.1257/jep.28.2.29

Bianco, D. (2012). On international spillovers. doi: 10.1016/J.ECONLET.2012.05.026

Blades, D. & Meyer zu Schlochtern, J. (1998). How to represent capital in international comparisons of total factor productivity. In *Second meeting of the canberra group on capital stock statistics.* Château de la Muette, Paris. Retrieved from `https://www.oecd.org/sdd/na/2662347.pdf` (Agenda item : 6b, Document number : 15)

Bloom, N., Schankerman, M. & van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, *81*(4), 1347–1393. Retrieved from `http://www.jstor.org/stable/23524180`

Blundell, R., Griffith, R. & van Reenen, J. (1995). Dynamic count data models of technological innovation. *Economic Journal*, *105*(429), 333-44. Retrieved from `https://EconPapers.repec.org/RePEc:ecj:econjl:v:105:y:1995:i:429:p:333-44`

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, *60*(3), 567–596. Retrieved from `http://www.jstor.org/stable/2951584`

Chetverikov, D. (2016). On cross-validated lasso in high dimensions. *The Annals of Statistics*. doi: 10.1214/20-aos2000

Cobb, C. W. & Douglas, P. H. (1928). A theory of production. *The American Economic Review*, *18*(1), 139–165. Retrieved from `http://www.jstor.org/stable/1811556`

Doi, J. (2004). Technological spillovers and patterns of growth with sector-specific rd. doi: 10.1016/J.JMACRO.2004.07.001

Eurostat. (2023). *Eurostat database on research and development.* Retrieved from `https://ec.europa.eu/eurostat/web/main/data/database`

León-Ledesma, M. (2000). R&d spillovers and export performance: Evidence from the oecd countries. *International Trade*. doi: 10.2139/ssrn.256630

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, *58*(304), 993–1010. Retrieved from http://www.jstor.org/stable/2283327

Lumenga-Neso, O., Olarreaga, M. & Schiff, M. (2005). On ‘indirect‘ trade-related r&d spillovers. *European Economic Review*, *49*(7), 1785-1798. Retrieved from https://EconPapers.repec.org/RePEc:eee:eecrev:v:49:y:2005:i:7:p:1785-1798

Manresa, E. (2016). Estimating the structure of social interactions using panel data. Retrieved from https://www.dropbox.com/s/erc8gr4yo0favyx/Manresa_2016_final.pdf?dl=0

Moretti, E. (2019). The intellectual spoils of war? defense r&d, productivity and international spillovers. *Conflict Studies: Domestic Politics eJournal*. doi: 10.3386/w26483

Park, S.-Y. (2010). Classification prediction error estimation system of microarray for a comparison of resampling methods based on multi-layer perceptron. *The Journal of the Korean Institute of Information and Communication Engineering*. doi: 10.6109/JKIICE.2010.14.2.534

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved from http://www.jstor.org/stable/2346178

A Appendix

Table A1: Descriptive statistics of the four variables from Eurostat.

	min	**max**	**mean**	**standard deviation**
GDP	6602	3267160	532196	716771
GERD	32	99554	10050	16142
Capital	611	581352	85985	118433
Labor	887	44131	9890	10710

Note: The *GDP*, *GERD* and *Capital* variables are expressed in millions of euros and the *Labor* variable in thousands of persons. All values are rounded to the nearest integer.

Table A2: List of the two-letter country codes of all countries used in Section 5, in ISO 3166-1 abbriviation.

Country Code	Country Name
AT	Austria
BE	Belgium
BG	Bulgaria
CZ	Czech Republic
DE	Germany
DK	Denmark
EL	Greece
ES	Spain
FI	Finland
FR	France
HU	Hungary
IE	Ireland
IT	Italy
LT	Lithuania
NL	Netherlands
PL	Poland
PT	Portugal
RO	Romania
SE	Sweden
SI	Slovenia
SK	Slovakia
UK	United Kingdom

Table A3: The elasticities of aggregate output with respect to the knowledge of country j: $\varepsilon^{Y}_{K_j}$.

	$\varepsilon^{Y}_{K_j}$
AT	6.8e-03
BE	-2.5e-04
BG	2.7e-05
CZ	-8.1e-04
DE	-1.6e-02
DK	1.9e-04
EL	-3.5e-03
ES	2.2e-03
FI	3.1e-03
FR	2.0e-02
HU	-1.5e-03
IE	-7.1e-04
IT	-4.1e-04
LT	2.8e-04
NL	3.7e-03
PL	-1.2e-03
PT	5.2e-05
RO	-5.6e-04
SE	7.5e-04
SI	-8.7e-04
SK	9.5e-04
UK	1.2e-03
ALL	0.01328

Note: The countries are abbriviated by two-letter country codes defined by ISO 3166-1. See Table A2 in the Appendix for their respective country names.

Table A4: The estimated spillover matrix on the international R&D spillovers using the fixed formula for the penalty term.

	AT	**BE**	**BG**	**CZ**	**DE**	**DK**	**EL**	**ES**	**FI**	**FR**	**HU**	**IE**	**IT**	**LT**	**NL**	**PL**	**PT**	**RO**	**SE**	**SI**	**SK**	**UK**
AT	0.077	0	0	0	0	0	0	0	0	0	0	0	0	0.057	0	0	0	0	0	0	0	0
BE	0	0.054	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BG	0	0	0.21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CZ	0	0	0	0.27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DE	0	0	0	0	0.075	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.3	0
DK	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EL	0	0	0	0	0	0	0.028	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ES	0	0	0	0	0	0	0	0.0015	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FI	0	0	0	0	0	0	0	0	0.096	0	0	0	0	0	0	0	0	0	0	0	0	0
FR	0	0	0	0	0	0	0	0	0	0.064	0	0	0	0	0	0	0	0	0	0	0	0
HU	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	0	0	0	0	0	0	0	0
IE	0	0	0	0	0	0	0	0	0	0	0	-0.019	0	0	0	0	0	0	0	0	0	0
IT	0	0	0	0	0	0	0	0	0	0	0	0	-0.025	0	0	0	0	0	0	0	0	0
LT	0	0	0	0	0	0	0	0	0	0	0	0	0	0.31	0	0	0	0	0	0	0	0
NL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.16	0	0	0.93	0	0	0	0
PL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33	0	0	0	0	0	0
PT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.11	0	0	0	0	0
RO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.28	0	0	0.26	0
SE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.11	0	0	0
SI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.16	0	0
SK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.26	0
UK	0	0	0	0	0	0	0	0	0	0	0	0	0	0.14	0	0	0	0	0	0	0	0.1

Note: The countries are abbriviated by two-letter country codes defined by ISO 3166-1. See Table A2 in the Appendix for their respective country names.

B Programming code

B.1 Simulation code

For the simulation code, we define two functions for estimating spillover effects in a simulated dataset. The first function, *HAC_weights*, simulates the data and uses weights that account for heteroskedasticity and autocorrelation to estimate the spillover effects. It generates a spillover effects matrix based on individuals randomly assigned to SIC codes and assigns spillover effects between individuals based on their SIC code similarity. It then fits a LASSO model iteratively for each individual using updated weights and calculates various statistics to evaluate the estimation accuracy. The second function, *no_weights*, also simulates the data but estimates the spillover effects without using weights. It generates the spillover effects matrix based on the same procedure as before and directly estimates the effects using the LASSO model. The two functions are put to use for all the results in Table 1 and 2.

B.2 European Data Application

In the coding for the empirical application, we prepare a dataframe by reading several CSV files and performing data manipulation and cleaning operations. We select specific columns from each dataframe, renames them, and joins them together based on common columns. We filter the dataset to include only observations from 1995 to 2017 and fills missing values with NA. Then we apply a custom function to remove countries that have gaps of at least two consecutive NA values. We then replace NA values with approximate values and adds logarithmic variants of the columns, and create lagged variables. We define the same two functions as described in the simulation coding above, for fitting LASSO models iteratively and calculating post-pooled LASSO estimates. The code then iterates over the data and fits LASSO models for each individual ID, computes residuals, and estimates coefficients using Frisch-Waugh-Lovell theorem. Finally, it calculates performance measures such as the elasticity of aggregate output with respect to knowledge and stores them in a dataframe.