IMT School for Advanced Studies, Lucca

Lucca, Italy

At the intersection between Machine Learning and Econometrics: theory and applications

PhD Program in "Systems Science"

Track in "Economics Networks and Business Analytics"

XXXIII Cycle

By

Federico Nutarelli

2021

The dissertation of Federico Nutarelli is approved.

PhD Program Coordinator: Prof. Dr. Rocco De Nicola, IMT School for Advanced Studies Lucca

Advisor: Prof. Massimo Riccaboni, IMT School for Advanced Studies Lucca

Co-Advisor: Prof. Giorgio Stefano Gnecco, IMT School for Advanced Studies Lucca

The dissertation of Federico Nutarelli has been reviewed by:

Prof. Ming Li, Zhejiang Normal University

Prof. Laura Magazzini, Sant'Anna School of Advanced Studies

Prof. Francesco Laio, Politecnico di Torino

IMT School for Advanced Studies Lucca
2021

To my parents, Paola and Giuseppe and my brother Marco

Contents

List of Figures

List of Tables

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Professor Massimo Riccaboni, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. Moreover I would like to thank my co-advisor, the one and only Giorgio Gnecco with whom we spent hours and hours trying to figure out solutions to complex problems. Thanks for giving me your support, answering my questions promptly and having thought the way to love math.

I would also like to thank my colleagues and friends who shared the PhD journey with me, Tatiana, Francesco, Dimitris, Emiliano, Matteo and Samuel. A big thanks goes also to seniors, Falco and Pietro, who opened me the way and were always there for discussing, advising me and stimulating me.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my friends, Alessandro and Gabriele, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

We are grateful to Crisis Lab. for the PHID database. We thank Prof. Jeff Wooldridge for the kind support in the application and interpretation of his innovative methodology. We would also like to thank FDA for having provided data on recalls and having cleared doubt on recalls' timing and procedures. We are grateful to Evaluate data for having allowed very up-to-date and detailed checks on clinical trials data. Finally, we are thankful to Springer AdInsight for having enabled a detailed classification at ATC 3 level of rare clinical trials.

Vita [1]

October 15, 1993	Born, Pietrasanta (Lucca), Italy
2017-	PhD Candidate Scholarship winner at IMT Alti Studi Lucca
2020-	Teaching Assistant at IMT Lucca for Advanced Studies in the course: "Applied Data Science"
2021-	Winner of the research fellowship "Analisi dell'innovazione in ambito biotecnologico e di valorizzazione delle terapie avanzate" (Analysis of innovation in the biotechnological context and in the enhancement of advanced therapies) in Roche Italy
2020-	Project Member 2020-2021 Italian project "Trade-off between Number of Examples and Precision in Variations of the Fixed-Effects Panel Data Model", funded by the Italian National Institute of High Mathematics (INdAM-GNAMPA). Principal Investigator: G. Gnecco
2015-2017	Master's Degree in Economics. Master Thesis in "Numerical methods for Economics: theories and applications" Final mark: 110/110 cum laude University of Pisa - Sant'Anna School for Advanced Studies, Pisa
2015	B.Sc. in Economics and Business Final mark: 110/110 cum laude University of Pisa, Pisa

[1] Please find the complete CV at `https://www.imtlucca.it/it/federico.nutarelli`

Academic Appointments

2021	Presented the conference paper "Optimal data collection design in machine learning: the case of the fixed effects generalized least squares panel data model" as a speaker at *Association for Mathematics applied to Social and Economic Sciences (AMASES2021)*
2021	Presented the conference paper "On dispersion curve identification" as a speaker at *Optimization and decision science (ODS2021)*
2019	Presented the conference paper "On the Trade-Off Between Number of Examples and Precision of Supervision in Regression" as a speaker at *The Fifth International Conference on Machine Learning, Optimization, and Data Science (LOD2019)*
2019	Summer School SIDE Summer School in "Machine Learning Algorithms for Econometricians" in Bertinoro.
2018	Summer School Summer School at TU Wien in "Machine Learning Methods and Data Analytics in Risk and Insurance" (VISS2018).

Additional Education and Awards

2016	Successful candidate of traineeship in ISTAT (Italian Institute of Statistics).

Publications

1. G. Gnecco, F. Nutarelli (2021). On the trade-off between number of examples and precision of supervision in machine learning problems. **Optimization Letters**, 15, 1711-1733
2. F. Nutarelli, F. Saracco, G. Caldarelli, A. Galli, A. Facchini (2019). "Network Analysis of the Mediterranean Food Footprint flows", Conference Paper **CCS2019** number 286
3. G. Gnecco, F. Nutarelli (2019). On the trade-off between number of examples and precision of supervision in regression problems. In Proc. of the 4th Int. Conf. of the International Neural Network Society on Big Data and Deep Learning **(INNS BDDL 2019)**, Springer, 1–6
4. G.Gnecco and F. Nutarelli (2020). Optimal trade-off between sample size and precision of supervision for the fixed effects panel data model. In Proc. of the 5^{th} Int. Conf. on machine Learning, Optimization & Data science **(LOD 2019), Lecture Notes in Computer Science**, 11943, 531–542
5. G. Gnecco, F. Nutarelli, D. Selvi (2021). Optimal data collection design in machine learning: the case of the fixed effects generalized least squares panel data model. **Machine Learning**, 110, 1549-1584
6. G. Gnecco, F. Nutarelli, D. Selvi (2020). Optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model. **Soft Computing**, 24, 15937–15949

Working Papers

1. F. Sforzi, A. Reale, F. Nutarelli, C. Drago. "Testing the quality of the delineation of Local Labour Market Areas", Working Paper
2. F. Nutarelli, A. Reale. "Commuting Flows in Tuscany: a general analysis", Working Paper
3. F. Nutarelli, M. Riccaboni. “Estimating own and cross price elasticities of drugs”, Working Paper.
4. A. Morescalchi, F. Nutarelli, M. Riccaboni. “Product recalls, market size and innovation in the pharmaceutical industry”, Working Paper.
5. G. Gnecco, F. Nutarelli, M. Riccaboni. “Matrix Completion of World Trade”, Working Paper.

Presentations

1. ASSA 2022 annual meeting 7-9 January `https://www.aeaweb.org/conference/`.
2. YETI, Young Economists of Tuscan Institutions (2021)
3. The Fifth International Conference on Machine Learning, Optimization, and Data Science (LOD 2019).
4. IMT Symposium 2019.
5. Reading group of "Machine Learning and Networks" IMT 2018-2020.
6. 4th Int. Conf. of the International Neural Network Society on Big Data and Deep Learning (INNS BDDL 2019), Sestri Levante, Italy, April 16th-18th, 2019.

Abstract

In the present work, we introduce theoretical and application novelties at the intersection between machine learning and econometrics in social and health sciences. In particular, Part 1 delves into optimizing the data collection process in a specific statistical model, commonly used in econometrics, employing an optimization criterion inspired by machine learning, namely, the generalization error conditioned on the training input data. In the first Chapter, we analyze and optimize the trade-off between sample size, the precision of supervision on a variation of the unbalanced fixed effects panel data model. In the second Chapter we extend the analysis to the Fixed Effects GLS (FEGLS) case in order to account for the heterogeneity in the data associated with different units, for which correlated measurement errors corrupt distinct observations related to the same unit. In Part 2, we introduce applications of innovative econometrics and machine learning techniques. In the third Chapter we propose a novel methodology to explore the effect of market size on market innovation in the Pharmaceutical industry. Finally, in the fourth Chapter, we innovate the literature on the economic complexity of countries through machine learning. The Dissertation contributes to the literature on machine learning and applied econometrics mainly by: (i) extending the current framework to novel scenarios and applications (Chapter 1 - Chapter 2); (ii) developing a novel econometric methodology to assess long-debated issues in literature (Chapter 3); (iii) constructing a novel index of economic complexity through machine learning (Chapter 4).

Introduction

Back in 2001, the Berkeley statistician Leo Breiman noted the presence of two distinct cultures in the use of statistical models (Breiman, 2001). In particular, he distinguished between a statistical approach based on a stochastic data generating process and one adopting algorithmic techniques and making no assumptions about the data generating process. When Breiman was writing, the two cultures were highly separated. Nowadays, the latter characterization does not apply, and the statistics community has largely accepted the Machine Learning (ML) revolution (as noted in Athey and Imbens, 2019 among others). This revolution has been facilitated by the exponentially grown ability of machines to solve complex tasks (Sejnowski, 2018). Yet many works and textbooks mix up econometrics and machine learning (see e.g., H. R. Varian, 2014a,Hastie, Tibshirani and Friedman, 2009 and Efron and Trevor Hastie, 2016).

Many empirical studies began to implement machine learning algorithms moved by the rapidly increasing methodological literature on the topic in recent years. More generally, though slower than in other fields, econometrics is moving away from the absolute dependence on pre-determined data models. A structural adjustment accompanies such a transition on how econometrics is perceived by academics as highlighted in Athey and Imbens (2019). Specifically, following Chamberlain (2000), if authors see econometrics as "decision making under uncertainty", their toolbox cannot ignore the fundamentals of machine learning.

Though relevant in general, machine learning mainly succeeded in "Big Data" contexts where a large amount of data per unit are available. Recently, machine learning techniques have been refined to also deal with the so-called panel-data structures, which provide information not only on the cross-sectional dimension but also on the

temporal one (i.e., units observed in time). The latter effort has been a further push toward the employment of ML techniques in econometrics settings, where panel data analysis constitutes a significant achievement. As reported in (Athey and Imbens, 2019) "for such setups, ML tools are becoming the standard across disciplines, and so the economist's toolkit needs to adapt accordingly while preserving the traditional strengths of applied econometrics".

The described environment opened the way to the application of ML also in social and health sciences. Starting from (H. R. Varian, 2014a) many authors in economics began to apply ML methodologies either to deepen and review research questions already explored in the past literature (Davis and Heller, 2017, Dubé and Misra, 2017,Thach, Kreinovich, and Trung, 2021, Y. Liu and Xie, 2019) or to raise new questions (Kleinberg, Lakkaraju, et al., 2018, Plonsky et al., 2017, Dash et al., 2019).

Most machine learning methods can be divided in two main branches with respect to the task addressed (Mitchell et al., 1997): (i) *unsupervised* learning and (ii) *supervised* learning models. In particular, unsupervised learning refers to models that try to draw inference starting from unlabeled variables Y. Unsupervised learning is mostly employed for clustering and pattern recognition. On the other hand, supervised learning refers to models that draw inference from labeled outcome variables Y (either discrete or continuous). Supervised learning is, above all, used to make predictions on outcomes corresponding to examples that were not used in the training process of the model (this is the case, e.g., of random forests and support vector machines). Since ML is specifically designed for predictive purposes (Kleinberg, Ludwig, et al., 2015), it provides proper tools for decision-makers. Furthermore, as correctly highlighted in (Bargagli Stoffi, 2020), *"the non-parametric nature of machine learning algorithms makes them suited to uncover hidden relationships between the predictors and the outcome variable that traditional econometric approaches could overlook. Indeed, the latter models – e.g., ordinary least squares and logistic regression – are built assuming a set of restrictions on the model's functional form and on the statistical distributions of the errors to guarantee statistical properties such as estimator unbiasedness and consistency. Machine learning algorithms often relax those assumptions, and the functional form is dictated by the data at hand (the phrase "data-driven models," often used to refer to machine learning algorithms, highlights this feature). This*

characteristic makes machine learning algorithms more adaptive and inductive, therefore enabling more accurate predictions for future outcome realizations."

The challenge for current researchers is to combine the significant advantages provided by the two disciplines of ML and econometrics. Yet, as correctly suggested in (H. R. Varian, 2014a), while ML provides solid algorithms for prediction tasks, econometrics uses robust statistical methods for prediction, inference, and above all, causal modeling of economic relationships.

In order to understand the deep penetration of ML – and Artificial Intelligence (AI) in general– in economics papers, Harding et al. (2018) compared the usage of ML techniques against standard econometrics methods within works in economics. They divided the analysis between settings with standard data sets and settings with big data sets. The authors commented on the overtake of ML techniques over traditional methods when dealing with big data as follows: *"due to the prevalence of connected digital devices, observational data sets are now available that are much larger and of higher frequency than traditional surveys: so-called Big Data. This has created opportunities for economists and policymakers to learn about economic systems and choices with a higher degree of precision. However, new methods, particularly those related to machine learning, are needed to take full advantage of Big Data. Furthermore, policymakers should consider a broader range of data as sensitive, researchers need checks to avoid unintentional bias, and economists should learn general-purpose coding languages.".*

The acquisition of big data, however, might be costly both in terms of time and money (Sivarajah et al., 2017). For instance, privacy issues on data acquisition make it hard to collect information (Lyko, Nitzschke, and Ngomo, 2016) properly. Besides, a significant issue that decision-makers might face is to achieve a suitable trad e-off between higher level of granularity – which implies a higher cost of collection and storage – and a lower level of granularity –which may lead to a small variety of queries that can be answered (Almeida, 2017). In particular, each example might have an acquisition cost. The latter cost, though often given, might be somehow controlled under certain circumstances. Yet, keeping the total acquisition cost (i.e., the one associated with the whole dataset) as fixed, one can intervene on the acquisition cost of each example and, thus, on their number.

The first Part of the Dissertation tries to partly solve the latter impediments by suggesting a way to optimize the trade-off between the precision of supervision and the number of examples. In line with (H. R. Varian, 2014a), future works on the topic will adapt the analysis to a causal context. The second Part displays the advantages of both disciplines of econometrics and ML through their application to social and health sciences. Before getting into the Dissertation core, let us set the stage on ML and its applications in econometrics contexts. Section 1 presents a very general overview of ML and econometrics, introducing the current discussion on how and when ML algorithms are beneficial to econometrics (and vice versa). Section 2 briefly reviews the literature on ML and econometrics. In Section 3, we draw the Dissertation outline and highlight how each Chapter contributes to the related literature.

1 Machine learning for econometrics

In his notorious paper about the novel tools provided by computer science (and in particular ML) to econometricians, Hal Varian highlighted how statistical and econometric analyses have to manage, nowadays, big datasets. Manipulating a large quantity of data, however, may raise specific issues: *"first, the sheer size of the data involved may require more powerful data manipulation tools. Second, we may have more potential predictors than appropriate for estimation, so we need to make some variable selection. Third, large datasets may allow for more flexible relationships than simple linear models. Machine learning techniques such as decision trees, support vector machines, neural nets, and deep learning, among others, may allow for more effective ways to model complex relationships."* (H. Varian, 2014).

Data analysis in statistics and econometrics can be divided into four parts according to H. R. Varian (2014a), which are prediction, summarization, estimation, and hypothesis testing. Machine learning can help with the first one by detecting hidden patterns in the data. Machine learning learns directly from data, without the restrictions (and assumptions) of model-based statistical methods (Breiman, 2001).

As an example, regarding the second issue, the traditional econometric analysis commonly performs summarization through linear regression. Usually, however, when it

comes to nonlinear relationships among data, conventional statistical tools encounter difficulties. Machine learning, instead, provides user-friendly algorithms quickly capturing nonlinear patterns (e.g., regression trees).
If, on the one hand, econometrics should learn from ML, on the other hand, it is also important that ML learns from econometrics (H. R. Varian, 2014a). In other words, authors should not abuse machine learning tools (see, e.g., Boelaert and Ollion, 2018 on the topic) but should rather smartly combine them with econometric techniques. The main reasons can be summarized as follows:

1. while machine learning is mainly about prediction (Wu et al., 2008), causal inference is about causality (see, e.g., Semykina and Wooldridge, 2010, Angrist and Pischke, 2008 among others), which is crucial for decision-makers.

2. there is still very few formed machine learning theory (above all theoretical Statistical Learning Theory) on how to deal with not Independently Identically Distributed (IID) data such as panel data and time series. No unified theory on the topic has been developed.

3. though research is making progress (see ,e.g., Athey, J. Tibshirani, and Wager, 2019), combining machine learning and causality is still an open debate.

As a critical example of reason 3, we mention here the problem of instrumentation. Instrumentation is a commonly employed method to correct for the endogeneity of one or more explanatory variables, say $X \in \mathbb{R}$. In particular, X is endogenous when correlated with unobservables, encoded in the error term ε. This condition violates the usual assumptions of traditional models such as Ordinary Least Squares (OLS) and requires the use of an instrument, i.e., a variable Z indirectly related to the outcome $Y \in \mathbb{R}$ through X and such that Z and ε are not correlated. Double Machine Learning has recently addressed instrumentation in the field of machine learning (Chernozhukov et al., 2017). Though capable of incorporating the advantages of machine learning (e.g., variable selection and dimensionality reduction), the latter technique does not provide a solid statistical theory for what concerns, for instance, instrumentation within a fixed-effect panel data model. More up-to-date tools to deal with endogeneity have been provided within the field of econometrics (W. Lin and Wooldridge, 2019) and, for this reason, they have been adopted throughout the

Discussion.
Therefore, it is vital to take advantage of machine learning in econometric contexts but, at the same time, to be aware of the goals of econometrics. Yet, as highlighted by Athey, J. Tibshirani, and Wager (2019) in their work about generalized random forests [2], applying machine learning to causality issues is a matter of moving the goal-posts. Significantly, the first step to take in order to combine machine learning and econometrics successfully is to select the model that primarily addresses the research question one would like to answer. Choosing the correct machine learning model is crucial since performance, computational scalability (i.e., the ability of the algorithm to handle more significant amounts of data), and interpretability differ across implementations. Usually, more complex algorithms require discretionary fine-tuning of hyper-parameters and are yet not ensured to deliver the best performance (Kotthoff, 2016). Scholars are, thus, required to run several algorithms and pick the best one in balancing interpretability and performance on the task that is studied.

Generally, in statistical prediction problems we are interested in understanding the conditional distribution of a variable Y as a function of some other variables, often referred to as *regressors*, $X = (x_1, \ldots, x_P)$. The $X-$variables are often called *features* or *controls* in machine learning. As stated in H. R. Varian (2014a), *"the focus of machine learning is to find some function that provides a good prediction of Y as a function of X"*. Typically, the authors are given some observed Y and X instances and want a "good" prediction of Y given new values of X. A prediction is considered as "good" if it minimizes some loss function such as the Mean Squared Error (MSE). We are interested in minimizing the loss *"associated with new out-of-sample observations"* (H. R. Varian, 2014a) of the X variable and not the observations used in the fit model. The optimal functional form minimizing the loss function is often modelled in two phases (Mullainathan and Spiess, 2017):

- conditional on a given level of complexity, pick the best in-sample loss-minimizing

[2] A technique that combines causality with random forests.

function $f(.)$ (Mullainathan and Spiess, 2017):

$$argmin \sum_{i=1}^{N} L(f(x_i), y_i) \quad over \quad f(.) \in F \quad s.t. \quad R(f(.)) \leq c,$$

where $\sum_{i=1}^{N} L(f(x_i), y_i)$ is the total in-sample loss to be minimized, $f(x_i)$ represents the predicted (or, adopting an econometric terminology, "fitted") values, y_i are the actual values, F is the class of admissible functions for the machine learning optimization problem, and R is a complexity functional, bounded to be lower than a suitably-chosen parameter $c \in \mathbb{R}$;

- estimate the optimal level of complexity using empirical tuning through cross-validation.

Cross-validation is a resampling procedure used to evaluate the performance of machine learning models on a limited data sample. In k-fold cross-validation (the most popular), data are shuffled randomly and divided into k groups. Repeatedly, one of the k groups is chosen as test set (i.e., to perform *out-of-sample* estimates), while the remaining $k-1$ groups are adopted as training set (i.e., to fit the model). The model is fit on the training set and evaluated on the test set. The model skill is measured by the average *out-of-sample* score. Since, by construction, machine learning algorithms are expected to perform exceptionally well on the training data, cross-validation is needed to avoid *overfitting*, i.e., finding a model closely mimicking or exactly replicating a specific set of data, which might, hence, fail to fit additional testing data or foresee future observations reliably. The root causes of overfitting are detailed in Ying (2019).

Under the described standard setup, econometricians may exploit the potentiality offered by machine learning algorithms. These include nonlinear methodologies as, for instance, classification/regression trees (CART), regression forests, matrix completion but also methods performing dimensionality reductions (e.g., LASSO and elastic nets).

Since, however, the flexibility of machine learning often comes at the expense of the un-verifiability of the underlying model assumptions (Bargagli-Stoffi, Niederreiter, and Riccaboni, 2021), econometricians must be careful in the adoption of simple

off-the-shelf algorithms Bargagli Stoffi (2020) from the ML literature. Indeed, ML techniques often require an intelligent adaptation to an econometrician's problem. As underlined by (Athey and Imbens, 2019), the most important adaptation *"is to exploit the structure of the problems, e.g., the causal nature of many estimands, the endogeneity of variables, the configuration of data such as panel data, the nature of discrete choice among a set of substitutable products, or the presence of credible restrictions motivated by economic theory"* as well as *"changing the optimization criteria of machine learning algorithms to prioritize considerations from causal inference"*. Further difficulties rely on the diversities in the terminology adopted in machine learning and econometrics.

Due to the complications mentioned above, a stream of literature is rapidly forming to instruct economists on manipulating machine learning algorithms and uniform the statistical theories embedded in the two disciplines of econometrics and machine learning. The effort of this newly born discipline is to insert in the mechanism of learning also some a-priori knowledge derived from other preliminary econometrics studies.

Reviewing the works about the adoption of machine learning in econometrics in detail goes beyond the scope of this Introduction, which, however, explains the main reasons why the following Dissertation tries to balance the adoption of econometrics and machine learning techniques and to partly fill the theoretical gaps detailed in Section 2. Tab.1 briefly outlines the major machine learning and econometrics methodologies employed.

Finally, Tab.2 points out the connections between the Chapters of the current Dissertation.

Table 1: Machine Learning and Econometrics techniques adopted in the Dissertation

Method	Description	Used in Chapter(s)
Fixed Effect GLS	Fixed Effects Generalised Least Squares (FEGLS) is a linear regression model with applications in several fields, able to represent unobserved heterogeneity in the data associated with different units (Wooldridge, 2002), for which distinct observations related to the same unit are corrupted by correlated measurement errors. We can also set-up the conditional variance of the output variable having the associated unit and input variables, by changing the cost per supervision of each training example	**1,2**
Fixed Effect IV	Fixed Effect Instrumental Variable (FEIV) is an innovative regression model that allows to control for the endogeneity of covariates. Differently from previous models, FEIV permits controlling for both idiosyncratic endogeneity and heterogeneity endogeneity. The technique is composed of two stages. A simple Wald test on the residuals' coefficient in the second stage allows verifying the presence of idiosyncratic endogeneity (W. Lin and Wooldridge, 2019).	**3**
Matrix Completion	Matrix Completion (MC) is a machine learning algorithm trying to fill in the missing entries of a partially observed matrix. Specifically, MC tries to minimize the trade-off (weighted sum) between a data-fitting term related to the known entries of the matrix and a regularization term, which avoids trivial reconstructions. A further regularization term avoids trivial binary reconstructions.	4
ROC, AUC, BACC	These are all techniques to measure the performance of classification algorithms in machine learning. Often, they are relative to a confusion matrix from where common evaluation metrics, such as true positive rate (TPR), true negative rate (TNR), and accuracy (ACC), can be easily calculated (Consoli, Reforgiato Recupero, and Saisana, 2021). ROC is a graph with TPR and false positive rate on its axes. AUC is the area under the ROC. BACC is an index adopted when observed data have unbalanced classifications.	4

2 Literature on machine learning for econometrics

The previous Section highlighted how machine learning provides valuable tools to econometricians, who should always check the models' underlying assumptions. Authors, moreover, should carefully select the best fitting machine learning algorithm to their specific research question, turning to advanced econometrics techniques whenever machine learning does not provide a solid statistical theory to its algorithms. The previous Section's take-home message is that econometricians should use machine learning without abusing it and having clear what is the exact question they would like to answer. As an example of the importance of the research question in determining whether to employ econometric or machine learning methodologies, suppose we have data on sales and advertising. Estimating the change in sales associated with the change in advertising expenditure *everything else held constant (ceteris paribus)* would raise a causality question (for which econometrics usually does a better job than machine learning). Instead, predicting the change in sales, one would

Table 2: Connections between the Chapteers presented in the Dissertation

Chapter	Connections
Chapters 1,2	The first two Chapters investigates a theoretical framework with methodologies. Its aim is to provide a theoretical framework where the remaining Chapters insert. Specifically, the first two Chapters introduce to the intersection between the disciplines of Machine Learning and Econometrics.
Chapter 3	The third Chapter provides an example of how, recent Econometrics tools are still powerful to overcome long debated issue in the literature. The instruments introduced in the first two Chapters are partially adopted to prove the latter statement.
Chapter 4	The last Chapter shows how Machine Learning is able to add to economic intuition and might provide a valid tool to support Econometrics. Thus, Chapter 4 validates the theses exposed in the previous Chapters.

expect to observe *when advertising expenditure changes (mutatis mutandis)* would raise a prediction question (for which machine learning usually works better).

For the reasons above and because machine learning algorithms represent a relatively new set of tools, combining machine learning and econometric methodologies to solve economics problems is still an open-debated topic. For instance, while it is true that there are many policy applications in which causality is not central, or even necessary (Kleinberg, Ludwig, et al., 2015), there are many others in which causality is key (see e.g. Hoover, 2008, Angrist and Pischke, 2008, Baltagi, 2006 among others). For example, suppose that we have a great machine learning model of the relationship between crime and police (H. R. Varian, 2014a). Such a model will not answer questions like "what happens to crime if we add further police." Indeed, the data generating process to answer the latter question is different from generating the crime-police relationship. In other words, the observed data are generated by a "more crime $\longrightarrow$ more police" rule, but we want to investigate what happens to crime when adding more police.

Due to the importance of the topic, a dedicated stream of literature on the applications of machine learning to econometrics was formed. To provide as much comprehensive as possible overview of the works on the subject, we will mainly rely on three works: H. R. Varian (2014a), Mullainathan and Spiess (2017), and Athey and Imbens (2019). Athey and Imbens (2019) categorize the literature on machine learning and econometrics according to their goals, methods, and settings. The following discussion will first characterize machine learning and econometrics goals, investigating how such goals have been reconciled in the literature. The works of Mullainathan and Spiess (2017) and H. R. Varian (2014a) will then help to understand the concrete tools that emerged from the mentioned reconciliation. In particular, we will exploit the studies

of Mullainathan and Spiess (2017) and H. R. Varian (2014a) to review the papers on the most used machine learning methods in econometrics contexts. A further distinction of the latter methods will be made according to their settings.

Athey and Imbens (2019) well summarize the goals of econometrics and machine learning. In particular, econometrics is mainly concerned with the estimation of a target being a functional of a joint distribution of data (William, 2008). The aim is often the estimation of a parameter –denoted as β– of a model describing the distribution of a set of variables in terms of a set of parameters. The focus of econometrics is on the quality of the estimator, usually quantified with significant sample efficiency. Typically, the interest also lies in building reliable confidence intervals and reporting the standard error of the estimates.

In contrast ML focus is on algorithms whose goal is to make prediction (Wu et al., 2008). The general example in Athey and Imbens (2019), clearly explains the different objectives of machine learning and econometrics adopting a common language to the two disciplines. Suppose that we want to model the conditional distribution of an outcome variable Y_i conditional on a regressor X_i:

$$Y_i|X_i \sim \mathcal{N}(\alpha + \beta^T X_i, \sigma^2),$$

where ^{T} indicates the transposition, $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 (Hellwig and K. M. Schmidt, 2002). One way to estimate $\theta = (\alpha, \beta)$ is through least squares:

$$(\hat{\alpha}_{ls}, \hat{\beta}_{ls}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} (Y_i - \alpha - \beta^T X_i)^2.$$

Most econometricians would adopt this model. In fact, if the data generating process is correct, the model has the known desirable asymptotic properties (e.g., the estimator is unbiased and efficient). In ML, however the goal is to predict Y_{N+1} for a new unit $N+1$ given X_{N+1}. The general problem is, therefore, minimizing the following loss

$$\mathbb{E}[Y_{N+1} - \hat{Y}_{N+1}]^2.$$

$\hat{Y}$ represents the prediction obtained with ML.

As stated in (Athey and Imbens, 2019), *"the question now is to come up with estimators* $(\hat{\alpha}, \hat{\beta})$ *that have good properties associated with this loss function. This need not be the least-squares estimator. In fact, when the dimension of the features exceeds two, we know from decision theory that we can do better in terms of expected squared error than the least-squares estimator. The latter is not admissible, i.e., other estimators dominate the least-squares estimator."* To dig into the latter mentioned issue, a stream of literature explores what Mullainathan and Spiess (2017) refer to as "prediction in the Service of Estimation." Authors working in this field try to review econometrics with the eyes of machine learning and vice versa. A part of this stream of literature dedicates to inference tasks having a prediction problem implicit inside them. As an example, overfitting can be described, from the viewpoint of econometrics, as an overcapacity of the prediction of $\hat{X}$ in two stages least squares procedures. In other words, following this analogy, overfitting happens when $\hat{X}$ picks up not only the part of variability explained by the instrument Z but also the noise. Under this light, overfitting becomes largely a problem of finite-sample bias, which can be solved through either regularization or empirical tuning. Regularization has already been introduced in several works. The papers exploit regularization within the first stage in a high-dimensional setting, including the LASSO (Belloni, Chernozhukov, and Wang, 2011) and ridge regression (C. Hansen and Kozbur, 2014). Recent contributions include nonlinearities in the functional forms (Hartford et al., 2016). Other tasks in causal inference also hide a prediction problem inside them. For instance –as mentioned in Mullainathan and Spiess (2017)– B. K. Lee, Lessler, and Stuart (2010) employ machine learning for constructing propensity scores, Chernozhukov et al. (2017) for managing high-dimensional controls in estimation of treatment effect and Imai and Strauss (2011), Athey, J. Tibshirani, and Wager (2019) (among others) for estimating heterogeneous causal effects. As clearly explained in Mullainathan and Spiess (2017), on the one hand, *"expressing parts of these inference tasks as prediction problems also makes clear the limitation on their output"*, on the other hand, one must be careful in interpreting the results (above all in tree representations) as they may suffer from instabilities (see the example on the American Housing Survey data in Mullainathan and Spiess, 2017).

Following the characterization of Athey and Imbens (2019), another stream of literature at the intersection between machine learning and econometrics is forming within settings with scarce information. Specifically, settings were discussed therein, where one observes information on many units (either as an outcome or as features). In many cases, however, econometrics data sets are equipped with partial information on the units one would like to analyze. A famous example of this setting is the so-called Netflix Problem (Bennett, Lanning, et al., 2007) where the aim was to impute the tastes of clients that never watched a movie based on the ratings given by similar users to that movie. To smartly solve such imputation issues, machine learning provides a novel technique called matrix completion. Matrix completion has recently been modified to fit panel data frameworks (see e.g. Candès and Recht, 2009, Mazumder, Trevor Hastie, and R. Tibshirani, 2010a and Athey, Bayati, et al., 2021). Though very useful in practice, matrix completion techniques have been –at the time being– only employed either as recommender systems for future purchases of consumers or for catching clients' preferences. One of the present work aims is to fill this gap in applied machine learning by extending matrix completion techniques to other fields of economics.

Finally, in this Dissertation, novel efforts have been made, in the field of theoretical machine learning for econometrics, in order to resolve the trade-off between datasets with high granularity but having higher collection costs and lower levels of granularity (with lower collection costs but less precise)[3]. Tab. 2 depicts the most up-to-date studies in economics, computer science, statistics, and engineering that deal with the last mentioned issue.

We conclude this review by quoting Mullainathan and Spiess (2017) on the importance of the figure of the economists in adopting machine learning algorithms. It is critical to remember that, though a powerful tool, machine learning is not always available for solving every econometric problem (as we will see in Chapter 3). Following Mullainathan and Spiess (2017) *"economists can play a crucial role in solving prediction policy problems. First, even though the prediction is important, machine learning is not enough: familiar econometric challenges arise. In deciding whether an algorithm*

[3]In the former case fewer examples are assuming that the total acquisition cost is the same in both circumstances.

could improve on the judge, one must resolve a basic counterfactual issue: we only know the crimes committed by those released. Many predictions problems share the feature that existing decision rules dictate the available data, and insights from the causal inference could prove helpful in tackling these problems; for example, Kleinberg et al. (2017) use pseudo-random assignment to judges of differing leniency in their application. Second, behavioral issues arise. Even when an algorithm can help, we must understand the factors that determine the adoption of these tools (Dawes, Faust, and Meehl 1989; Dietvorst, Simmons, and Massey 2015; Yeomans, Shah, Mullainathan, and Kleinberg 2016). What factors determine faith in the algorithm? Would a simpler algorithm be more believed? How do we encourage judges to use their private information optimally? These questions combine problems of technology diffusion, information economics, and behavioral economics".

Table 3: Abbreviations: STAT = statistics; ENG = engineering; ML-EC = machine learning - econometrics

Author, year	**Domain**	**Type of analysis**
Groves et al. (2004a)	STAT	optimal survey design
Nguyen et al. (2008a)	ENG	optimal design of measurement devices
Gnecco, Nutarelli, and Selvi (2020)	**ML-EC**	**a priori optimization and optimal data collection design in machine learning: the case of unbalanced F.E.**
C.-H. Chen and L. H. Lee (2011a)	STAT	Optimal Computing Budget Allocation (OCBA); a posteriori

3 Outline of the Dissertation

The present Dissertation is made of two parts. In Part 1, we explore the field of theoretical machine learning for econometrics by introducing a novel technique to optimize the trade-off between training set size and precision of supervision in a panel data context. Specifically, we carry out the last effort by extending the interaction between machine learning and optimization to econometrics. An introductory Chapter will present the basic model in Gnecco and Nutarelli (2019c) consisting of the optimization of the mentioned trade-off in a balanced panel F.E. context. In the first Chapter we extend the model in Gnecco and Nutarelli (2019c) by analyzing and optimizing the trade-off between sample size, the precision of supervision (the reciprocal of the conditional variance of the output) and selection probabilities in an unbalanced fixed effects context. We assume an upper bound on the expected total supervision cost and fix the expected number of observed units for each instant (Gnecco, Nutarelli, and Selvi, 2020). In the second Chapter we propose a further extension of the model introduced in Gnecco and Nutarelli (2019c). Specifically, we relax the assumption of independent measurement errors made in the previous model. Furthermore, in the former work, the trade-off between training set size and precision of supervision was analyzed only for a fixed number of units, assuming that the number of observations associated with the same unit is large enough to justify a large-sample approximation with respect to the number of observations. In Gnecco, Nutarelli, and Selvi (2021) we provide a large-sample approximation with respect to the number of units and a large-sample approximation with respect to both the number of units and the number of observations per unit. Other differences will be detailed throughout the Discussion. In Part 2, we introduce applications of machine learning and econometrics. Specifically, following Mullainathan and Spiess (2017), in the third Chapter we present a problem in which we believe that econometrics provides more robust tools than machine learning. Finally, in the fourth Chapter we show how, according to Athey and Imbens (2019) and H. R. Varian (2014a), machine learning and economic intuition can cooperate, leading to interesting new perspectives. Specifically, we applied, for the first time, Matrix Completion techniques to develop a novel index of economic complexity.

The present Dissertation contributes to the literature mainly by (i) extending the current framework in the literature of theoretical machine learning to novel applications and scenarios related to econometrics (Chapter 1 - Chapter 2); (ii) applying novel econometric and machine learning methodologies to solve open-debated questions in the literature (Chapter 3 - Chapter 4); (iii) rethinking economic issues using machine learning tools (Chapter 4); (iv) extending the interaction of machine learning and optimization also to econometrics (Chapter 1 - Chapter 2).
We introduce below a more profound overview of the mentioned works by specifying the main research questions (R.Q.s) they tackle.

Part 1: Machine learning for econometrics

Chapter 1: optimal trade-off between the number of examples and precision of supervision in the F.E. unbalanced panel case

The first Chapter is focused on the unbalanced fixed effects panel data model. This is a linear regression model representing unobserved heterogeneity in the data by allowing each two distinct observational unit to have possibly different numbers of associated observations. We specifically address the case in which the model includes the additional possibility of controlling the conditional variance of the output given the input and the selection probabilities of the different units per unit time. This is achieved by varying the cost associated with the supervision of each training example. Assuming an upper bound on the expected total supervision cost and fixing the expected number of observed units for each time instant, we analyze and optimize the trade-off between sample size, the precision of supervision (the reciprocal of the conditional variance of the output) and selection probabilities. This is obtained by formulating and solving a practical optimization problem. The formulation of such a problem is based on a large-sample upper bound on the generalization error associated with the estimates of the parameters of the unbalanced fixed effects panel data model, conditioned on the training input dataset. We prove that, under appropriate assumptions, in some cases, "many but bad" examples provide a smaller large-sample upper bound on the conditional generalization error than "few but good" ones, whereas in other cases, the opposite occurs. To summarize, the research question of this Chap-

ter focuses on the optimal trade-off between sample size, the precision of supervision, and selection probabilities for a general linear model of the input-output relationship, which is the unbalanced fixed effects panel data model. Loosely speaking,

RQ1: is having "many but bad" examples always worse –in terms of the minimization of the generalization error– than having "few but good" examples in an unbalanced fixed effects context?

Chapter 2: optimal trade-off between the number of examples and precision of supervision: the case of correlated measurement errors in a F.E. scenario.

This Chapter and the previous one belong to the strand of literature that combines machine learning, optimization, and econometrics. The aim is to optimize the data collection process in a specific statistical model, commonly used in econometrics, employing an optimization criterion inspired by machine learning, namely, the generalization error conditioned on the training input data. More specifically, the Chapter is focused on the analysis of the conditional generalization error of the Fixed Effects Generalized Least Squares (FEGLS) panel data model, i.e., a linear regression model with applications in several fields, able to represent unobserved heterogeneity in the data associated with different units, for which correlated measurement errors corrupt distinct observations related to the same unit. Hence, the research question of this Chapter is similar to that of the previous one, but considers the case of correlated measurement errors. In particular,

RQ2: is having "many but bad" examples always worse –in terms of the minimization of the generalization error– than having "few but good" examples in a fixed-effects context when the measurement errors are correlated? How can we compare the two cases?

This analysis, like the one of the first Chapter, is fascinating in an era where it is critical to understand which data will provide the significant benefits to organizations and people in general (Günther et al., 2017).

The framework considered in this Chapter differs from the classical FEGLS model for

the additional possibility of controlling the conditional variance of the output variable given the associated unit and input variables by changing the cost per supervision of each training example. Assuming an upper bound on the total supervision cost, i.e., the cost associated with the whole training set, the trade-off between the training set size and the precision of supervision (i.e., the reciprocal of the conditional variance of the output variable) is analyzed and optimized. This is achieved by formulating and solving in closed form suitable optimization problems based on large-sample approximations of the generalization error associated with the FEGLS estimates of the model parameters, conditioned on the training input data. The results of the analysis extend to the FEGLS case and to various large-sample approximations of its conditional generalization error, the ones obtained in the previous work (Gnecco and Nutarelli, 2019a) for simpler linear regression models. They highlight the importance of how the precision of supervision scales with respect to the cost per training example in determining the optimal trade-off between training set size and precision. Numerical results confirm the validity of the theoretical findings.

The power of econometrics: the relationship between size and innovation

Though recent algorithms in machine learning have been proposed for dealing with causality (Athey, J. Tibshirani, and Wager, 2019), they still do not offer robust tools that are key to such kinds of analyses. As mentioned in the Introduction, most of these tools developed by econometricians are concerned with panel data structures and endogeneity. This Chapter introduces a well-established problem of reverse causality (endogeneity) within a panel context in the economic literature. In such a scenario, econometrics provides advanced models specifically designed to deal with the presented technical issues (W. Lin and Wooldridge, 2019). In the following, we will give a general overview of the Chapter.

The responsiveness of investments in research to market rewards is recognized in the literature on markets for innovation (Jacob, 1966;Acemoglu and Linn, 2004; Dear et al., 2021). The empirical evidence suggests that a change in market size, like the one measured by demographical shifts, is associated with an increase in the number of new available drugs (Acemoglu and Linn, 2004; Dubois et al., 2015). Nonetheless, there is still an open debate about potential reverse causality (Cerda, 2007). In this

Chapter, we investigate the effect of market size on innovation as measured by active clinical trials.

RQ3: How does market size impact on market innovation?

We exploit product recalls, an innovative instrument. We test that the latter is sharp, strong, and unexpected to market. The work analyses the relationship between US market size and innovation at the ATC-3 level. The adopted dataset is original and granular (ATC-4 level). We exploit a novel two-step IV methodology proposed by W. Lin and Wooldridge (2019). The results reveal a robust and significantly positive response of the number of active trials to market size.

The combination of machine learning and economic intuition: an application of matrix completion to economic complexity

Chapter 4 concludes the Dissertation by showing how economic intuition can boost machine learning and vice versa. Specifically, in contrast with Chapter 3, we display here a scenario in which machine learning can help dealing with a traditional economic issue, namely economic complexity of countries.
This Chapter combines Matrix Completion (MC) – a class of machine-learning methods commonly used in the context of recommendation systems – with a recently developed economic complexity index (GENEPY) by Sciarra, Chiarotti, Ridolfi, et al. (2020a), to compare the observed and expected economic complexity of countries. While the former is computed by a direct application of GENEPY to the incidence matrix associated with the Revealed Comparative Advantage (RCA) matrix, the latter is based on using MC to reconstruct a discretized version of the RCA matrix, then applying GENEPY to the resulting surrogate incidence matrix. According to our knowledge, this is the first time MC is used in combination with an economic complexity index. As a by-product of the analysis, it is shown that the false positive rate per country of a binary classifier constructed starting from the average entry-wise output of MC can be used as a proxy of GENEPY. Finally, a novel Matrix cOmpletion iNdex of Economic complexitY (MONEY), based on MC, is introduced, which

is related to the predictability of the entries of the RCA matrix associated with each country (the lower the predictability, the higher the complexity), and also considers the product dimension.
The last research question is, thus,

RQ4: can ML help in developing a novel index of economic complexity?

As evident from this brief overview, the common denominator of the present Dissertation is combining machine learning and econometrics to produce novel evidence. Future research will be devoted to relating the evidence to real-world scenarios (Part 1) and extending the outcomes of the analyses to novel setups (Part 2).

Part 1: Machine Learning for Econometrics

Chapter 1

Optimal data collection design in machine learning: the case of unbalanced Fixed Effects [1]

The content of the present Chapter comes from the following sources: the baseline model is taken directly from Gnecco, Nutarelli, and Selvi (2020), while the unbalanced Fixed Effects model refers to Gnecco, Nutarelli, and Selvi (2020).

1 Baseline model

1.1 Introduction

In several applications in economics, engineering, and many other fields, one has to approximate a function from a finite set of input-output noisy examples. This belongs to the typical class of problems studied by supervised machine learning (Vapnik,

[1] This chapter is partially based on Gnecco, G., Nutarelli, F. & Selvi, D. Optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model. Soft Comput 24, 15937–15949 (2020). https://doi.org/10.1007/s00500-020-05317-5

1998). In some cases, the noise variance of the output can be reduced to some extent by the researcher, by increasing the cost of each supervision. For instance, measurement devices with larger precision (hence, larger cost) could be employed. Similarly, the supervision could be provided by teachers (e.g., doctors) with a higher level of expertise (hence, higher cost). In all such cases, the investigation of an optimal trade-off between the sample size and the precision of supervision is needed. In Gnecco and Nutarelli (2019b), this analysis was conducted employing a modification of the classical linear regression model, in which one is additionally given the possibility of controlling the conditional variance of the output given the input, by varying the time (hence, the cost) dedicated to the supervision of each training example, and fixing an upper bound on the total available supervision time. Based on a large-sample approximation of the output of the ordinary least squares regression algorithm, it was shown therein that the optimal choice of the supervision time per example is highly dependent on the noise model.

In this preliminary section to Section 2 and Section 3, we draw from Gnecco and Nutarelli (2019c), which consider a general linear model of the input-output relationship, i.e. the fixed effects panel data model. In this model, observations associated with different units (individuals) are associated with different constants, which are able to represent unobserved heterogeneity in the data. Moreover, the same unit is observed along another dimension, which is typically time. The model is commonly applied in the analysis of microeconomics and macroeconomics data (Wooldridge, 2010), where each unit may represent, e.g., a firm, or a country. It is also applied in biostatistics (Frees et al., 2004), psychology, political science, sociology, and life sciences (Andreß, Golsch, and A. W. Schmidt, 2013). Though the aim of this preliminary section is only to provide a clear background for Section 2 and Section 3, it is important to provide an overview of the results found in such a simplified scenario. Indeed, they will be similar to the ones recovered in later sections. Specifically, consistently with the results of the analysis performed in Gnecco and Nutarelli (2019b), for the classical linear regression model, we show – in Gnecco and Nutarelli (2019c)– that, also for the fixed effects panel data model, the following holds: when the precision of the supervision (the reciprocal of the conditional variance of the output) increases less than proportionally with respect to the supervision cost per example,

the minimum (large-sample approximation of the) generalization error (conditioned on the training input data) is obtained in correspondence of the smallest supervision cost per example (hence, of the largest number of examples); when the precision increases more than proportionally with respect to the supervision cost per example, the optimal supervision cost per example is the largest one (which is associated with the smallest number of examples). In summary, the results of the theoretical analyses performed in Gnecco and Nutarelli (2019b) and, for a different regression model, in this section highlight that an increase of the sample size is not always beneficial, if it is feasible to collect a smaller number of more reliable data. Hence, not only their number, but also their quality matters. This looks particularly relevant when one has the possibility of designing the data collection process.

1.2 Background of baseline model

We recall some basic facts about the following (static) fixed effects panel data model (see, e.g., Wooldridge, 2010, Chapter 10)[2]:

$$y_{n,t} := \eta_n + \boldsymbol{\beta}' \boldsymbol{x}_{n,t}, n = 1, \ldots, N, t = 1, \ldots, T\,. \tag{1.1}$$

Here, the outputs $y_{n,t}$'s are scalar, whereas the inputs $\boldsymbol{x}_{n,t}$'s ($n = 1, \ldots, N, t = 1, \ldots, T$) are column vectors in $\mathbb{R}^p$, which are modeled as random vectors. The parameters of the model are the individual constants η_n ($n = 1, \ldots, N$), one for each unit, and the column vector $\boldsymbol{\beta} \in \mathbb{R}^p$. Eq. (1.1) represents a balanced panel data model, in which each unit n is associated with the same number T of outputs, each one at a different time t.

Noisy measures $\tilde{y}_{n,t}$'s of the outputs $y_{n,t}$'s are available. They are generated according to the following additive noise model:

$$\tilde{y}_{n,t} := y_{n,t} + \varepsilon_{n,t}\,, \tag{1.2}$$

where the $\varepsilon_{n,t}$'s are mutually independent and identically distributed random variables, having mean 0 and the same variance σ^2. Moreover, they are independent also

[2]For simplicity of exposition, here the model is not presented in its most general form (e.g., the disturbances $\varepsilon_{n,t}$'s are simply assumed to be mutually independent).

from all the $\boldsymbol{x}_{n,t}$'s.

The training input-output pairs $(\boldsymbol{x}_{n,t}, \tilde{y}_{n,t})$ $(n = 1, \ldots, N, t = 1, \ldots, T)$ are used to estimate the parameters of the model. Assuming the invertibility of the matrix $\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q} \boldsymbol{X}_n$, the fixed effects estimate of $\boldsymbol{\beta}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{FE} &:= \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q} \boldsymbol{X}_n \right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q} \tilde{\boldsymbol{y}}_n \right) \\ &= \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}' \boldsymbol{Q} \boldsymbol{X}_n \right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}' \boldsymbol{Q} \tilde{\boldsymbol{y}}_n \right), \end{aligned} \tag{1.3}$$

where $\boldsymbol{X}_n \in \mathbb{R}^{T,p}$ is a matrix whose rows are the transposes of the $\boldsymbol{x}_{n,t}$'s, $\tilde{\boldsymbol{y}}_n$ is a column vector which collects the noisy measures $\tilde{y}_{n,t}$'s,

$$\boldsymbol{Q} := \boldsymbol{I}_T - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T' \tag{1.4}$$

(being $\boldsymbol{I}_T \in \mathbb{R}^{T \times T}$ the identity matrix, and $\mathbf{1}_T \in \mathbb{R}^T$ a column vector whose elements are all equal to 1) is a symmetric and idempotent matrix (i.e., $\boldsymbol{Q}' = \boldsymbol{Q} = \boldsymbol{Q}^2$). Hence, for each unit n,

$$\boldsymbol{Q} \boldsymbol{X}_n = \begin{bmatrix} \boldsymbol{x}_{n,1} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_{n,t} \\ \boldsymbol{x}_{n,2} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_{n,t} \\ \cdots \\ \boldsymbol{x}_{n,T} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_{n,t} \end{bmatrix}, \tag{1.5}$$

and

$$\boldsymbol{Q} \boldsymbol{y}_n = \begin{bmatrix} y_{n,1} - \frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{n,t} \\ y_{n,2} - \frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{n,t} \\ \cdots \\ y_{n,T} - \frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{n,t} \end{bmatrix} \tag{1.6}$$

represent, respectively, the matrix of time de-meaned training inputs, and the vector of time de-meaned corrupted training outputs.

The fixed effects estimates of the η_n's are

$$\hat{\eta}_{n,FE} := \frac{1}{T} \sum_{t=1}^{T} \left(\tilde{y}_{n,t} - \hat{\boldsymbol{\beta}}_{FE}' \boldsymbol{x}_{n,t} \right). \tag{1.7}$$

The estimates (1.30) and (1.31) are unbiased, i.e.,

$$\mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right\}=\mathbf{0}_p \tag{1.8}$$

(being $\mathbf{0}_p \in \mathbb{R}^p$ a column vector whose elements are all equal to 0), and

$$\mathbb{E}\left\{\hat{\eta}_{n,FE}-\eta_n\right\}=0\,. \tag{1.9}$$

Finally, the covariance matrix of $\hat{\boldsymbol{\beta}}_{FE}$, conditioned on the training input data $\{\boldsymbol{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}$, is

$$\begin{aligned}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{FE}|\{\boldsymbol{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right) &= \sigma^2\left(\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}\boldsymbol{X}_n\right)^{-1} \\ &= \sigma^2\left(\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_n\right)^{-1}. \end{aligned} \tag{1.10}$$

1.3 Conditional generalization error and its large-sample approximation in the baseline model

We define the generalization error for the i-th unit ($i=1,\ldots,N$), conditioned on the training input data, as follows:

$$\mathbb{E}\left\{\left(\hat{\eta}_{i,FE}+\hat{\boldsymbol{\beta}}_{FE}'\boldsymbol{x}_i^{test}-\eta_i-\boldsymbol{\beta}'\boldsymbol{x}_i^{test}\right)^2\Big|\{\boldsymbol{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\}, \tag{1.11}$$

where $\boldsymbol{x}_i^{test}\in\mathbb{R}^p$ is independent from the training data. It is the expected mean squared error of the prediction of the output associated with a test input, conditioned on the training input data.

For $n=1,\ldots,N$, let $\boldsymbol{\varepsilon}_n\in\mathbb{R}^T$ be the column vector whose elements are the $\varepsilon_{n,t}$'s, and let $\boldsymbol{\eta}_n\in\mathbb{R}^T$ be the column vector whose elements are all equal to η_n. Using

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_m'\right\}=0 \tag{1.12}$$

for $n\neq m$,

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n'\right\}=\sigma^2\boldsymbol{I}_T\,, \tag{1.13}$$

$$\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}\,, \tag{1.14}$$

$$\boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}'\boldsymbol{Q}\,, \tag{1.15}$$

$$\boldsymbol{Q}\boldsymbol{\eta}_n = \boldsymbol{Q}'\boldsymbol{\eta}_n = \boldsymbol{0}_T\,, \tag{1.16}$$

and

$$\boldsymbol{Q}\boldsymbol{1}_T = \boldsymbol{Q}'\boldsymbol{1}_T = \boldsymbol{0}_T\,, \tag{1.17}$$

we can simplify the expression (1.32) of the conditional generalization error as follows, highlighting its dependence on σ^2 and T (see the Appendix for the details):

$$\begin{aligned}(1.32) \;=\; & \frac{\sigma^2}{T^2}\boldsymbol{1}_T'\boldsymbol{X}_i\left(\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}_i'\boldsymbol{1}_T + \frac{\sigma^2}{T} \\ & + \mathbb{E}\left\{\sigma^2\left(\boldsymbol{x}_i^{test}\right)'\left(\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_i^{test}\Big|\left\{\boldsymbol{x}_{n,t}\right\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\ & - 2\mathbb{E}\left\{\frac{\sigma^2}{T}\boldsymbol{1}_T'\boldsymbol{X}_i\left(\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_i^{test}\Big|\left\{\boldsymbol{x}_{n,t}\right\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\}. \end{aligned} \tag{1.18}$$

Next, we obtain a large-sample approximation of the conditional generalization error (1.39) with respect to T, for a fixed number of units N. Such an approximation is useful, e.g., in the application of the model to macroeconomics data[3], for which it is common to investigate the case of a large horizon T.

Under mild conditions[4], the following convergences in probability[5] hold, which follow from Chebyschev's law of large numbers Ruud et al., 2000, Section 13.4.2:

$$\text{plim}_{T\to+\infty}\frac{1}{T}\boldsymbol{1}_T'\boldsymbol{X}_i = \left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\right)'\,, \tag{1.19}$$

[3]The case of finite T and large N is of more interest for microeconometrics, and will be investigated in future research.

[4]E.g., if the $\boldsymbol{x}_{n,t}$'s are independent, identically distributed, and have finite moments up to the order 4.

[5]We recall that a sequence of random real matrices $\boldsymbol{M}_T$, $T = 1,\ldots,+\infty$ converges in probability to the real matrix $\boldsymbol{M}$ if, for every $\varepsilon > 0$, $\text{Prob}\left(\|\boldsymbol{M}_T - \boldsymbol{M}\| > \varepsilon\right)$ (where $\|\cdot\|$ is an arbitrary matrix norm) tends to 0 as T tends to $+\infty$. In this case, one writes $\text{plim}_{T\to+\infty}\boldsymbol{M}_T = \boldsymbol{M}$.

and

$$\text{plim}_{T\to+\infty}\frac{1}{T}\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_n=\boldsymbol{A}_N\,, \tag{1.20}$$

where

$$\boldsymbol{A}_N=\boldsymbol{A}_N':=\sum_{n=1}^{N}\mathbb{E}\left\{(\boldsymbol{x}_{n,1}-\mathbb{E}\left\{\boldsymbol{x}_{n,1}\right\})'(\boldsymbol{x}_{n,1}-\mathbb{E}\left\{\boldsymbol{x}_{n,1}\right\})\right\} \tag{1.21}$$

is a symmetric and positive semi-definite matrix. In the following, its positive definiteness (hence, its invertibility) is also assumed[6].

When (2.20) and (1.42) hold, the conditional generalization error (1.39) has the following large-sample approximation with respect to T[7]:

$$\begin{aligned}(1.39)\;\simeq\;&\frac{\sigma^2}{T}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\right)'\boldsymbol{A}_N^{-1}\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}+\frac{\sigma^2}{T}\\&+\frac{\sigma^2}{T}\mathbb{E}\left\{\left(\boldsymbol{x}_i^{test}\right)'\boldsymbol{A}_N^{-1}\boldsymbol{x}_i^{test}\right\}-2\frac{\sigma^2}{T}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\right)'\boldsymbol{A}_N^{-1}\mathbb{E}\left\{\boldsymbol{x}_i^{test}\right\}\\=\;&\frac{\sigma^2}{T}\left(1+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right),\end{aligned} \tag{1.22}$$

where $\|\cdot\|_2$ denotes the l_2-norm, and $\boldsymbol{A}_N^{-\frac{1}{2}}$ is the principal square root (i.e., the symmetric and positive definite square root) of the symmetric and positive definite matrix $\boldsymbol{A}_N^{-1}$.

Interestingly, the large-sample approximation (1.46) has the form $\frac{\sigma^2}{T}K_i$, where

$$K_i:=\left(1+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right) \tag{1.23}$$

is a positive constant. This simplifies the analysis of the trade-off between sample size and precision of supervision performed in Gnecco and Nutarelli (2019c), since one does not need to compute the exact expression of K_i to find the optimal trade-off.

[6]The existence of the probability limit (1.42) and the assumed positive definiteness of the matrix $\boldsymbol{A}_N$ guarantee that the invertibility of the matrix $\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}\boldsymbol{X}_n=\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_n$ (see Section 2) holds with probability near 1 for large T.

[7]This is obtained taking also into account that, as a consequence of the Continuous Mapping Theorem Florescu, 2014, Theorem 7.33, the probability limit of the product of two random variables equals the product of their probability limits, when the latter two exist.

Exploring in details the optimization of the trade-off of the set up presented in this section goes beyond the aims of this introductory section, which just wanted to formally introduce the topics discussed in more details in the next sections. The general aim of the following sections is to perform a similar analysis in more articulated (but useful in practice) econometrics context. The generalization error will change according to the specific set up. In order to give an intuition of the type of results that the optimization of generalization errors of the type described in Eq.(1.21) will output, we briefly summarize the outcomes of the optimization of this simple preliminary problem. In particular it can be proved that, by modeling in a suitable way the variance as a function of the supervision cost per example c [8], one can approximately minimize the conditional generalization error by solving the following optimization problem:

$$\text{minimize}_{c\in[c_{\min},c_{\max}]} N K_i k c^{1-\alpha}\,, \tag{1.24}$$

whose optimal solutions c° have the following expressions:

1. if $0<\alpha<1$ ("decreasing returns of scale"): $c^\circ = c_{\min}$;

2. if $\alpha>1$ ("increasing returns of scale"): $c^\circ = c_{\max}$;

3. if $\alpha=1$ ("constant returns of scale"): $c^\circ =$ any cost c in the interval $[c_{\min}, c_{\max}]$.

In summary, the results of the analysis show that, in the case of "decreasing returns of scale", "many but bad" examples are associated with a smaller generalization error than "few but good" ones. The opposite occurs for "increasing returns of scale", whereas the case of "constant returns of scale" is intermediate. These results are qualitatively in line with the ones obtained in Gnecco and Nutarelli (2019b) for a simpler linear regression problem, to which the ordinary least squares algorithm was applied. This depends on the fact that, in both cases, the conditional generalization error has the functional form $\frac{\sigma^2}{T}K_i$ (although two different positive constants K_i are involved in the two different cases).

[8] Further details will follow in later sections.

2 Optimal data collection design in machine learning: the case of unbalanced fixed effects

In this Section, we analyze the optimal trade-off between sample size, precision of supervision, and selection probabilities for a more general – compared to the baseline model in Section 1 – linear model of the input-output relationship, which is the unbalanced fixed effects panel data model. As aforementioned, the (either balanced or unbalanced) fixed effects model is commonly applied in the econometric analysis of microeconomic and macroeconomic data (Andreß, Golsch, and A. W. Schmidt, 2013; Arellano, 2003; Cameron and Trivedi, 2005; Wooldridge, 2010), where each unit may represent, e.g., a firm, or a country. It is also applied in other fields (see the discussion in Section 1). In Section 1 we introduced the fixed effect panel data context only formally. Here, we try to provide a brief qualitative description before proceeding with the main analysis.

In a fixed effects panel data model, observations related to different observational units (individuals) are associated with possibly different constants, which are able to represent unobserved heterogeneity in the data. Moreover, the same unit is observed along another dimension, which is typically time. In the unbalanced case, at each instant, different units may be not observed with some positive probability (possibly unit-dependent), resulting in a possibly unbalanced panel.In this framework, the balanced case corresponds to the situation in which the number of observations is the same for all the units.

The present set up extends significantly the analysis of Section 1 to the unbalanced fixed effects panel data model, which is more general than the balanced case considered therein, and leads to an optimization problem that is more complex to investigate. In fact, in Section 1, all the units are always selected at each instant [9], therefore the selection probabilities do not appear as optimization variables in the corresponding model. Moreover, theoretical arguments are reported in much more details in the current section.

The results that will be presented in this Section concerning the unbalanced fixed effects panel data model are consistent with those of Section 1 for the balanced case,

[9] In other words, every unit appears at time t,

and those of Gnecco and Nutarelli (2019b) and Gnecco and Nutarelli (2019c) concerning simpler linear regression models. Specifically, we show that, also for the unbalanced fixed effects panel data model, the following holds. When the precision of the supervision increases less than proportionally with respect to the supervision cost per example, the minimum (large-sample upper bound on the) generalization error (conditioned on the training input dataset) is obtained in correspondence of the smallest supervision cost per example. As a consequence of the problem formulation, this corresponds to the choice of the largest number of examples. Instead, when the precision of the supervision increases more than proportionally with respect to the supervision cost per example, the optimal supervision cost per example is the largest one. Again, as a consequence of the problem formulation, this corresponds to the choice of the smallest number of examples. The structure of the optimal selection probabilities is also investigated, under the constraint of a constant expected number of observed units for each instant. In summary, the results of the theoretical analyses performed, for different regression models of increasing complexity, in Section 1, Gnecco and Nutarelli (2019c), and in this Section highlight that, in some circumstances, collecting a smaller number of more reliable data is preferable than increasing the size of the sample set. This looks particularly relevant when one is given a certain flexibility in designing the data collection process.

Up to our knowledge, the analysis and the optimization of the trade-off between sample size, precision of supervision, and selection probabilities in regression has been carried out rarely in the machine-learning literature. Thus, the contribution given to the literature by this Section is mainly to provide novel applications of machine learning by combining the latter discipline with econometrics. Nevertheless, the approach applied in this Section resembles the one used in the optimization of sample survey design, where some of the design parameters are optimized to minimize the sampling variance (Groves et al., 2004b). Such an approach is also similar to the one exploited in Nguyen et al. (2008b) for the optimization of the design of measurement devices. In that framework, however, linear regression is marginally involved, since only arithmetic averages of measurement results are considered therein. The search for optimal sample designs can be also performed by the Optimal Computing Budget Allocation (OCBA) method (C.-H. Chen and L. H. Lee, 2011b). Differently

from that approach, however, our analysis provides the optimal design *a priori*, i.e., before actually collecting the data. The work of this Section can also be related to recent literature dealing with the joint application of machine learning, optimization, and econometrics (H. R. Varian, 2014b; Athey and Imbens, 2016; Stoffi and Gnecco, 2018; Stoffi and Gnecco, 2020; Andrew, 2017). For instance, the generalization error - which is typically investigated by machine learning, and optimized by solving suitable optimization problems - is not addressed in the classical analysis of the either balanced or unbalanced fixed effects panel data model (Wooldridge, 2010, Sections 10 and 17). Finally, an advantage of the approach considered in this Section with respect to other possible ones grounded on Statistical Learning Theory (SLT) (Vapnik, 1998) is that, being based on a large-sample approximation, it provides bounds on the conditional generalization error that do not need any a-posteriori evaluation of empirical risks.

The Section is structured as follows. Subsection 2 provides a background on the unbalanced fixed effects panel data model. Subsection 2 presents the analysis of its conditional generalization error, and of the large-sample upper bound on the latter with respect to time. Subsection 2 formulates and solves the optimization problem modeling the trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model, using the large-sample upper bound above. Finally, Subsection 2 discusses some possible applications and extensions of the theoretical results obtained in the work.

2.1 Background for the unbalanced case

The baseline model is the same as the one of Section 1, except that now the total time in which a unit is observed is unit dependent.

Specifically, in the (static) unbalanced fixed effects panel data model (see, e.g., Wooldridge, 2010, Sections 10 and 17) each observational unit $n = 1, \ldots, N$ is observed in the time instants $t = 1, \ldots, T_n$. As a consequence, with respect to the baseline model, the inputs $\boldsymbol{x}_{n,t}$ to the model will be again random column vectors in $\mathbb{R}^p$ but observed for $n = 1, \ldots, N$ and – here lies the main difference compared to Section 1 – for $t = 1, \ldots, T_n$. The same applies to the scalar output $y_{n,t} \in \mathbb{R}$. The individual constants η_n $(n = 1, \ldots, N)$ and the column vector $\boldsymbol{\beta} \in \mathbb{R}^p$ do not differ from those presented

in the baseline model since they are time independent. The (noise-free) input-output relationship is similar to that of Equation (1.1) but differs in that each unit is not observed for the same amount of time:

$$y_{n,t} := \eta_n + \boldsymbol{\beta}' \boldsymbol{x}_{n,t}, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T_n \,. \tag{1.25}$$

Equation (1.25) represents an unbalanced panel data model, which can be applied in the following two situations:

- distinct units n are associated with possibly different numbers T_n of data collected at each time instant $t = 1, \ldots, T_n$ over a whole observation period $T \geq \max_{n=1}^{N} T_n$;
- the observations related to the same unit are associated with a subsequence $\{t_1, t_2, \ldots, t_{T_n}\}$ of the sequence $\{1, 2, \ldots, T\}$.

In the next subsections, we focus on the second situation which is also the most likely in econometrics scenarios. To avoid burdening the notation by introducing an additional index, we still indicate, also in this case, by $\{1, 2, \ldots, T_n\}$ the subsequence $\{t_1, t_2, \ldots, t_{T_n}\}$. A possible way to get different numbers of observations T_n for distinct units consists in associating to each unit n a scalar $q_n \in (0, 1]$, which denotes the (positive) probability that n is observed at any time t. Selections for different units are supposed to be mutually independent. For simplicity, for each unit, selections at different times are also assumed to be mutually independent. For a total observation time T, denoting by $\mathbb{E}$ the expectation operator, the expected number of observations for each unit n is $\mathbb{E}\{T_n\} = q_n T$. The balanced case, which was considered in Section 1, corresponds to the situation $q_n = 1$ for each n.
The general notation of the present Section is very close to that in Section 1, but is adapted to accommodate the possibly unbalanced nature of the data. Below we report the modified notation for the unbalanced case. Often the only difference in the notation compared o Section 1 lies on the fact that T is substituted with T_n.
As usual, $\{\varepsilon_{n,t}\}_{n=1,\ldots,N,\, t=1,\ldots,T_n}$ represents the collection of mutually independent and identically distributed random variables, having mean 0 and the same variance σ^2. Moreover, let the $\varepsilon_{n,t}$ be independent also from all the $\boldsymbol{x}_{n,t}$. As in Section 1, noisy measurements $\tilde{y}_{n,t}$ of the outputs $y_{n,t}$ are available also in the unbalanced case; specif-

ically, the following additive noise model is considered:

$$\tilde{y}_{n,t} = y_{n,t} + \varepsilon_{n,t}, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T_n \,. \tag{1.26}$$

The input-output pairs $(\boldsymbol{x}_{n,t}, \tilde{y}_{n,t})$ for $n = 1, \ldots, N$, $t = 1, \ldots, T_n$, are used to train the model, i.e., to estimate its parameters. In the following, for $n = 1, \ldots, N$, let $\boldsymbol{X}_n \in \mathbb{R}^{T_n, p}$ denote the matrix whose rows are the transposes of the $\boldsymbol{x}_{n,t}$; $\tilde{\boldsymbol{y}}_n$ be the column vector that collects the noisy measurements $\tilde{y}_{n,t}$; $\boldsymbol{I}_{T_n} \in \mathbb{R}^{T_n \times T_n}$ denote the identity matrix; $\mathbf{1}_{T_n} \in \mathbb{R}^{T_n}$ be the column vector whose elements are all equal to 1; and

$$\boldsymbol{Q}_n := \boldsymbol{I}_{T_n} - \frac{1}{T_n} \mathbf{1}_{T_n} \mathbf{1}_{T_n'} \tag{1.27}$$

be a symmetric and idempotent matrix, i.e., such that $\boldsymbol{Q}_n' = \boldsymbol{Q}_n = \boldsymbol{Q}_n^2$. Hence –similarly to Section 1– for each unit n,

$$\boldsymbol{Q}_n \boldsymbol{X}_n = \begin{bmatrix} \boldsymbol{x}_{n,1} - \frac{1}{T_n} \sum_{t=1}^{T_n} \boldsymbol{x}_{n,t} \\ \boldsymbol{x}_{n,2} - \frac{1}{T_n} \sum_{t=1}^{T_n} \boldsymbol{x}_{n,t} \\ \cdots \\ \boldsymbol{x}_{n,T_n} - \frac{1}{T_n} \sum_{t=1}^{T_n} \boldsymbol{x}_{n,t} \end{bmatrix}, \tag{1.28}$$

and

$$\boldsymbol{Q}_n \tilde{\boldsymbol{y}}_n = \begin{bmatrix} \tilde{y}_{n,1} - \frac{1}{T_n} \sum_{t=1}^{T_n} \tilde{y}_{n,t} \\ \tilde{y}_{n,2} - \frac{1}{T_n} \sum_{t=1}^{T_n} \tilde{y}_{n,t} \\ \cdots \\ \tilde{y}_{n,T_n} - \frac{1}{T_n} \sum_{t=1}^{T_n} \tilde{y}_{n,t} \end{bmatrix} \tag{1.29}$$

represent, respectively, the matrix of time de-meaned training inputs, and the vector of time de-meaned corrupted training outputs. We remember here that the aim of time de-meaning is to generate another dataset that does not include the fixed effects, making it possible to estimate first the vector $\boldsymbol{\beta}$, then - going back to the original dataset - the fixed effects η_n.

Assuming in the following the invertibility of the matrix $\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n$ (see the next Remark 6 for a mild condition ensuring this), the fixed effects estimate of $\boldsymbol{\beta}$ for the

unbalanced case is

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_{FE} \\
:= \;& \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \tilde{\boldsymbol{y}}_n\right) \\
= \;& \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \tilde{\boldsymbol{y}}_n\right). \qquad (1.30)
\end{aligned}$$

The unbalanced Fixed Effects (FE) estimates of the η_n, for $n = 1, \ldots, N$, are

$$\hat{\eta}_{n,FE} := \frac{1}{T_n} \sum_{t=1}^{T_n} \left(\tilde{y}_{n,t} - \hat{\boldsymbol{\beta}}_{FE}' \boldsymbol{x}_{n,t}\right). \qquad (1.31)$$

Properties 1.8-1.10 hold also in the unbalanced context.

2.2 Large-sample upper bound on the conditional generalization error in the unbalanced framework

This section analyzes the generalization error associated with the FE estimates (1.30) and (1.31), conditioned on the training input dataset, by providing its large-sample approximation, and a related large-sample upper bound on it. Then, in the next section, the resulting expression is optimized, after choosing a suitable model for the variance σ^2 of the measurement noise, and imposing appropriate constraints. Though the expression of the large-sample approximation error looks similar to those of Section 1, the different dimension of its components plays an important role as we will see later on in the discussion. For the sake of completeness, we report below a complete description of the expression of the generalization error in the context of unbalanced fixed effects.

Let $\boldsymbol{x}_i^{test} \in \mathbb{R}^p$ be a random test vector, which is assumed to have finite mean and finite covariance matrix, and to be independent from the training data. We express the generalization error for the i-th unit ($i = 1, \ldots, N$), conditioned on the training

input dataset, as follows[10]:

$$\mathbb{E}\left\{\left(\hat{\eta}_{i,FE}+\hat{\boldsymbol{\beta}}'_{FE}\boldsymbol{x}_i^{test}-\eta_i-\boldsymbol{\beta}'\boldsymbol{x}_i^{test}\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}. \tag{1.32}$$

The conditional generalization error (1.32) represents the expected mean squared error of the prediction of the output associated with a test input, conditioned on the training input dataset.

For $n=1,\ldots,N$, let $\boldsymbol{\varepsilon}_n\in\mathbb{R}^{T_n}$ be the column vector whose elements are the $\varepsilon_{n,t}$; $\boldsymbol{\eta}_n\in\mathbb{R}^{T_n}$ be the column vector whose elements are all equal to η_n; and $\mathbf{0}_{T_n\times T_n}\in\mathbb{R}^{T_n\times T_n}$ be the matrix whose elements are all equal to 0. Noting that

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}'_m\right\}=\mathbf{0}_{T_n\times T_n},\quad \text{for } n\neq m, \tag{1.33}$$

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}'_n\right\}=\sigma^2\boldsymbol{I}_{T_n}, \tag{1.34}$$

$$\boldsymbol{Q}'_n\boldsymbol{Q}_n=\boldsymbol{Q}_n, \tag{1.35}$$

$$\boldsymbol{Q}'_n\boldsymbol{Q}_n\boldsymbol{Q}'_n\boldsymbol{Q}_n=\boldsymbol{Q}'_n\boldsymbol{Q}_n, \tag{1.36}$$

$$\boldsymbol{Q}_n\boldsymbol{\eta}_n=\boldsymbol{Q}'_n\boldsymbol{\eta}_n=\mathbf{0}_{T_n}, \tag{1.37}$$

and

$$\boldsymbol{Q}_n\mathbf{1}_{T_n}=\boldsymbol{Q}'_n\mathbf{1}_{T_n}=\mathbf{0}_{T_n}, \tag{1.38}$$

we can express the conditional generalization error (1.32) as follows, highlighting its dependence on σ^2 and T_i (see Appendix 1 for the details):

$$\begin{aligned}
&\mathbb{E}\left\{\left(\hat{\eta}_{i,FE}+\hat{\boldsymbol{\beta}}'_{FE}\boldsymbol{x}_i^{test}-\eta_i-\boldsymbol{\beta}'\boldsymbol{x}_i^{test}\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
=&\frac{\sigma^2}{T_i^2}\mathbf{1}'_{T_i}\boldsymbol{X}_i\left(\sum_{n=1}^N\boldsymbol{X}'_n\boldsymbol{Q}'_n\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}'_i\mathbf{1}_{T_i}\\
&+\frac{\sigma^2}{T_i}
\end{aligned}$$

[10]See the next Remark 3 for a justification of the choice of the conditioned generalization error for the analysis, instead of its unconditional version.

$$+\mathbb{E}\left\{\sigma^2\left(\boldsymbol{x}_i^{test}\right)'\left(\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_i^{test}\Big|\{\boldsymbol{X}_n\}_{n=1}^{N}\right\}$$
$$-2\mathbb{E}\left\{\frac{\sigma^2}{T_i}\mathbf{1}_{T_i}'\boldsymbol{X}_i\left(\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_i^{test}\Big|\{\boldsymbol{X}_n\}_{n=1}^{N}\right\}. \tag{1.39}$$

Next, we obtain a large-sample approximation of the conditional generalization error (1.39) with respect to T, for a fixed number N of units[11].

For $n=1,\ldots,N$, let the symmetric and positive semi-definite matrices $\boldsymbol{A}_n\in\mathbb{R}^{p\times p}$ be defined as

$$\boldsymbol{A}_n=\boldsymbol{A}_n':=\mathbb{E}\left\{\left(\boldsymbol{x}_{n,1}-\mathbb{E}\left\{\boldsymbol{x}_{n,1}\right\}\right)\left(\boldsymbol{x}_{n,1}-\mathbb{E}\left\{\boldsymbol{x}_{n,1}\right\}\right)'\right\}. \tag{1.40}$$

In the following, the positive definiteness (hence, the invertibility) of each matrix A_n is assumed. This is a quite mild condition because it is associated with the fact that, with probability 1, the random vectors $\boldsymbol{x}_{n,1}-\mathbb{E}\{\boldsymbol{x}_{n,1}\}$ do not belong to a subspace of $\mathbb{R}^p$ with dimension smaller than p (so, they are effectively p-dimensional random vectors).

Under mild conditions (e.g., if the $\boldsymbol{x}_{n,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4), the following convergences in probability[12] hold:

$$\begin{aligned}&\mathrm{plim}_{T\to+\infty}\frac{1}{T_i}\mathbf{1}_{T_i}'\boldsymbol{X}_i\\ =\;&\mathrm{plim}_{T\to+\infty}\frac{1}{T_i}\sum_{t=1}^{T_i}\boldsymbol{x}_{i,t}'\\ =\;&\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\right)',\end{aligned} \tag{1.41}$$

[11]Such an approximation is useful, e.g., in the application of the model to macroeconomics data, for which it is common to investigate the case of a large horizon T. The case of finite T and large N is of more interest for microeconometrics Cameron and Trivedi, 2005, and will be investigated in future research.

[12]We recall that a sequence of random real matrices $\boldsymbol{M}_T$ of the same dimension, $T=1,2,\ldots$, converges in probability to the real matrix $\boldsymbol{M}$ if, for every $\varepsilon>0$, $\mathrm{Prob}\left(\|\boldsymbol{M}_T-\boldsymbol{M}\|>\varepsilon\right)$ (where $\|\cdot\|$ is an arbitrary matrix norm) tends to 0 as T tends to $+\infty$. In this case, one writes $\mathrm{plim}_{T\to+\infty}\boldsymbol{M}_T=\boldsymbol{M}$.

and

$$\begin{aligned} & \text{plim}_{T\to+\infty}\frac{1}{T}\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n \\ = \; & \text{plim}_{T\to+\infty}\sum_{n=1}^{N}\frac{T_n}{T}\frac{1}{T_n}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n \\ = \; & \boldsymbol{A}_N\,, \end{aligned} \tag{1.42}$$

where

$$\boldsymbol{A}_N = \boldsymbol{A}_N' := \sum_{n=1}^{N} q_n\boldsymbol{A}_n \tag{1.43}$$

which is the weighted summation, with positive weights q_n, of the symmetric and positive definite matrices $\boldsymbol{A}_n$, hence it is also a symmetric and positive definite matrix.

Remark 1. *Equations (1.41) and (1.42) follow from the extension of Chebyschev's weak law of large numbers (Ruud et al., 2000, Section 13.4.2) to the case of the summation of a random number of mutually independent random variables (Révész, 1968, Theorem 10.1), combined with other technical results. First, for each* $n = 1,\ldots,N$*, convergence in probability of* $\frac{1}{T_n}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n$ *to* $\boldsymbol{A}_n$ *is proved element-wise, by applying Révész (1968, Theorem 10.1). Then, one exploits the fact that, as a consequence of the Continuous Mapping Theorem (Florescu, 2014, Theorem 7.33), the probability limit of the product of two random variables (in this case,* $\frac{T_n}{T}$ *and each element of* $\frac{1}{T_n}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n$*) equals the product of their probability limits, when the latter two exist (which is the case for* $\frac{T_n}{T}$ *and each element of* $\frac{1}{T_n}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n$*). Finally, one applies the fact that, for a random matrix, element-wise convergence in probability implies convergence in probability of the whole random matrix (Herriges, 2011).*

Remark 2. *The existence of the probability limit (1.42) and the positive definiteness of the matrix* $\boldsymbol{A}_N$ *guarantee that the invertibility of the matrix*

$$\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}_n\boldsymbol{X}_n = \sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n \tag{1.44}$$

(see Section 2) holds with probability close to 1 *for large* T*. Due to the generalization of Slutsky's theorem reported in Greene (2003, Theorem D.14)*[13]*, under the stated assumptions also the sequence of random matrices*

$$\left(\frac{1}{T}\sum_{n=1}^{N}\boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1} \tag{1.45}$$

converges in probability to $\boldsymbol{A}_N^{-1}$*. This is needed to obtain the next large-sample approximation (1.46) of the conditional generalization error.*

Remark 3. *We point out that the conditional generalization error (1.32) is investigated in this work, instead of its unconditional version because, in general, probability limits and expectations cannot be inverted in order. This could prevent the application of Greene (2003, Theorem D.14) (or of similar results about probability limits) when performing a similar analysis for the unconditional generalization error.*

Let $\|\cdot\|_2$ denote the l_2-norm, and $\boldsymbol{A}_N^{-\frac{1}{2}}$ be the principal square root (i.e., the symmetric and positive definite square root) of the symmetric and positive definite matrix $\boldsymbol{A}_N^{-1}$. When (1.41) and (1.42) hold, from (1.39) and the assumed independence of $\boldsymbol{x}_i^{test}$ from all the other random vectors we get the following large-sample approximation (with respect to T) for the conditional generalization error (1.32):

$$\begin{aligned}
& \mathbb{E}\left\{\left(\hat{\eta}_{i,FE}+\hat{\boldsymbol{\beta}}_{FE}'\boldsymbol{x}_i^{test}-\eta_i-\boldsymbol{\beta}'\boldsymbol{x}_i^{test}\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\} \\
\simeq\ & \frac{\sigma^2}{T}\left(\mathbb{E}\{\boldsymbol{x}_{i,1}\}\right)'\boldsymbol{A}_N^{-1}\mathbb{E}\{\boldsymbol{x}_{i,1}\} \\
& +\frac{\sigma^2}{q_iT} \\
& +\frac{\sigma^2}{T}\mathbb{E}\left\{\left(\boldsymbol{x}_i^{test}\right)'\boldsymbol{A}_N^{-1}\boldsymbol{x}_i^{test}\right\} \\
& -2\frac{\sigma^2}{T}\left(\mathbb{E}\{\boldsymbol{x}_{i,1}\}\right)'\boldsymbol{A}_N^{-1}\mathbb{E}\left\{\boldsymbol{x}_i^{test}\right\} \\
=\ & \frac{\sigma^2}{T}\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\{\boldsymbol{x}_{i,1}\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right).
\end{aligned} \tag{1.46}$$

[13]It states that, given a sequence of random real square matrices $\boldsymbol{M}_T$ of the same dimension, $T=1,2,\ldots$, if $\text{plim}_{T\to+\infty}\boldsymbol{M}_T=\boldsymbol{B}$ and $\boldsymbol{B}$ is invertible, then also $\text{plim}_{T\to+\infty}\boldsymbol{M}_T^{-1}=\boldsymbol{B}^{-1}$.

In the following, we denote, for a generic symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{s\times s}$, by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$, respectively, its minimum and maximum eigenvalue. Starting from the large-sample approximation (1.46), the following steps can be proved (see Appendix 1 for the details):

$$\begin{aligned}
&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\\
\leq&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\lambda_{\max}(\boldsymbol{A}_N^{-1})\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\\
=&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\frac{1}{\lambda_{\min}(\boldsymbol{A}_N)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\\
\leq&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\frac{1}{\sum_{n=1}^N q_n\lambda_{\min}(\boldsymbol{A}_n)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right).
\end{aligned} \tag{1.47}$$

We refer to the inequality

$$\begin{aligned}
&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\\
\leq&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\frac{1}{\sum_{n=1}^N q_n\lambda_{\min}(\boldsymbol{A}_n)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)
\end{aligned} \tag{1.48}$$

as the large-sample upper bound on the conditional generalization error. Interestingly, its right-hand side is expressed in the separable form $\frac{\sigma^2}{T}K_i(\{q_n\}_{n=1}^N)$, where

$$\begin{aligned}
&K_i(\{q_n\}_{n=1}^N)\\
:=&\left(\frac{1}{q_i}+\frac{1}{\sum_{n=1}^N q_n\lambda_{\min}(\boldsymbol{A}_n)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)
\end{aligned} \tag{1.49}$$

depends only on the q_n. As shown in the next section, this simplifies the analysis of the trade-off between sample size, precision of supervision, and selection probabilities performed therein, since one does not need to compute the exact expression of the

function $K_i(\{q_n\}_{n=1}^N)$ to find the optimal trade-off with respect to a suitable subset of optimization variables.

2.3 Optimal trade-off between sample size, precision of supervision, and selection probabilities

In this section, we are interested in optimizing the large-sample upper bound (1.48) of the conditional generalization error when the variance σ^2 is modeled as a decreasing function of the supervision cost per example c, and a given upper bound $C > 0$ is imposed on the expected total supervision cost $\sum_{n=1}^N q_n T c$ associated with the whole training set. For large T, this upper bound practically coincides with the total supervision cost $\sum_{n=1}^N T_n c$. This follows by an application of Chebyschev's weak law of large numbers.

Remark 4. *In our previous conference work Gnecco and Nutarelli (2019c), the large-sample approximation (1.46) was optimized, instead of (1.48). This was motivated by the fact that all the selection probabilities q_n were fixed to* 1*, implying that both q_i and A_N, hence also the term*

$$\left(\frac{1}{q_i} + \mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\{\boldsymbol{x}_{i,1}\} - \boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right), \tag{1.50}$$

were constant therein.

In the following analysis of the optimal trade-off, N is kept fixed; furthermore, one imposes the constraints

$$q_{n,\min} \leq q_n \leq q_{n,\max}, \quad n = 1, \ldots, N, \tag{1.51}$$

for some given $q_{n,\min} \in (0,1)$ and $q_{n,\max} \in [q_{n,\min}, 1]$, and

$$\sum_{n=1}^N q_n = \bar{q} N, \tag{1.52}$$

for some given $\bar{q} \in \left[\frac{\sum_{n=1}^N q_{n,\min}}{N}, \frac{\sum_{n=1}^N q_{n,\max}}{N}\right] \subseteq (0,1]$. In Equation (1.52), $\sum_{n=1}^N q_n$ repre-

sents the expected number of observed units for each instant, which is fixed. Moreover, T is chosen as $\left\lfloor \frac{C}{\bar{q}Nc} \right\rfloor$. Finally, the supervision cost per example c is allowed to take values on the interval $[c_{\min}, c_{\max}]$, where $0 < c_{\min} < c_{\max}$, so that the resulting T belongs to $\left\{ \left\lfloor \frac{C}{\bar{q}Nc_{\max}} \right\rfloor, \ldots, \left\lfloor \frac{C}{\bar{q}Nc_{\min}} \right\rfloor \right\}$. In the following, C is supposed to be sufficiently large, so that the large-sample upper bound (1.48) can be assumed to hold for every $c \in [c_{\min}, c_{\max}]$ and every $q_n \in [q_{n,\min}, q_{n,\max}]$ (for $n = 1, \ldots, N$).

Consistently with Section 1, Gnecco and Nutarelli (2019c), we adopt the following model for the variance σ^2, as a function of the supervision cost per example c:

$$\sigma^2(c) = kc^{-\alpha}, \tag{1.53}$$

where $k, \alpha > 0$. For $0 < \alpha < 1$, if one doubles the supervision cost per example c, then the precision $1/\sigma^2(c)$ (i.e., the reciprocal of the conditional variance of the output) becomes less than two times its initial value (or equivalently, the variance $\sigma^2(c)$ becomes more than one half its initial value). This case is referred to as "decreasing returns of scale" in the precision of each supervision. Conversely, for $\alpha > 1$, if one doubles the supervision cost per example c, then the precision $1/\sigma^2(c)$ becomes more than two times its initial value (or equivalently, the variance $\sigma^2(c)$ becomes less than one half its initial value). This case is referred to as "increasing returns of scale" in the precision of each supervision. Finally, the case $\alpha = 1$ is intermediate and refers to "constant returns of scale". The broad idea, here, can be to assume a function with a sigmoidal shape, modeling saturation effects of the precision to the cost ($p(c)$ being p the precision of supervision and c the related cost). The mentioned function is either concave or convex according to the range considered. The main factors leading to increasing/constant/decreasing return to scale are technology and know-how. In all the cases above, the precision of each supervision increases by increasing the supervision cost per example c.

Summarizing, under the assumptions above, the optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed

effects panel data model is modeled by the following optimization problem:

$$\begin{aligned} \underset{\substack{c\in[c_{\min},c_{\max}],\\ q_n\in[q_{n,\min},q_{n,\max}],\\ n=1,\dots,N}}{\text{minimize}} \quad & K_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor\frac{C}{\bar{q}Nc}\right\rfloor} \\ \text{s.t.} \quad & \sum_{n=1}^N q_n = \bar{q}N\,. \end{aligned} \tag{1.54}$$

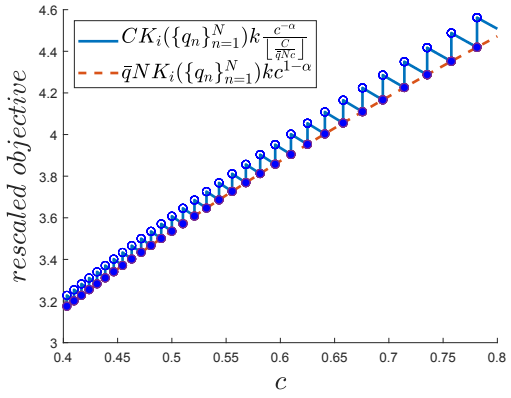
By a similar argument as in the proof of Gnecco and Nutarelli (2019c, Proposition 3.2), which refers to an analogous function approximation problem, when C is sufficiently large, the objective function $CK_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor\frac{C}{\bar{q}Nc}\right\rfloor}$ of the optimization problem (1.54), rescaled by the multiplicative factor C, can be approximated, with a negligible error in the maximum norm on $[c_{\min},c_{\max}]\times\Pi_{n=1}^N[q_{n,\min},q_{n,\max}]$, by $\bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha}$. Figure 2 shows the behavior of the rescaled objective functions

$$CK_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor\frac{C}{\bar{q}Nc}\right\rfloor} \tag{1.55}$$

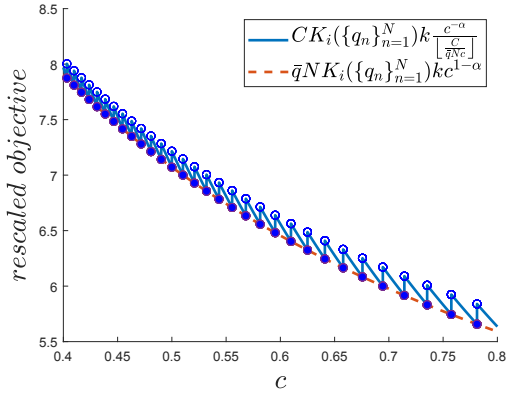
and

$$\bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha} \tag{1.56}$$

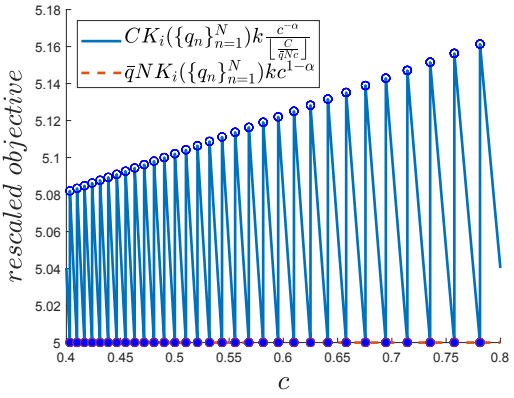
for the three cases $0<\alpha=0.5<1$, $\alpha=1.5>1$, and $\alpha=1$. The values of the other parameters are $k=0.5$, $\bar{q}=0.5$, $K_i(\{q_n\}_{n=1}^N)=2$ (which can be assumed to hold for a fixed choice of the set of the q_n), $N=10$, $C=125$, $c_{\min}=0.4$, and $c_{\max}=0.8$. One can show by standard calculus that, for $C\to+\infty$ and the q_n fixed to constant values, the number of discontinuity points of the rescaled objective function $CK_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor\frac{C}{\bar{q}Nc}\right\rfloor}$ tends to infinity, whereas the amplitude of its oscillations above the lower envelope $\bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha}$ tends to 0 uniformly with respect to $c\in[c_{\min},c_{\max}]$.
The latter result is similar to those obtained in Section 1 for the baseline scenario. Below, we provide a graphical representation of the rescaled objective functions.



(a)



(b)



(c)

Figure 1: Plots of the rescaled objective functions $CK_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\lfloor\frac{C}{\bar{q}Nc}\rfloor}$ and $\bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha}$ for $\alpha = 0.5$ (a), $\alpha = 1.5$ (b), and $\alpha = 1$ (c). The unbalanced panel data case.

Concluding, under the approximation above, one can replace the optimization problem (1.54) with

$$\begin{aligned} \underset{\substack{c\in[c_{\min},c_{\max}],\\ q_n\in[q_{n,\min},q_{n,\max}],\\ n=1,\dots,N}}{\text{minimize}} \quad & \bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha} \\ \text{s.t.} \quad & \sum_{n=1}^{N} q_n = \bar{q}N\,. \end{aligned} \tag{1.57}$$

Such optimization problem appears in a separable form, in which one can optimize separately the variable c and the variables q_n, for $n = 1, \dots, N$. In particular, the optimal solutions c° have the following expressions:

a) if $0 < \alpha < 1$ ("decreasing returns of scale"): $c^\circ = c_{\min}$;

b) if $\alpha > 1$ ("increasing returns of scale"): $c^\circ = c_{\max}$;

c) if $\alpha = 1$ ("constant returns of scale"): $c^\circ =$ any cost c in the interval $[c_{\min}, c_{\max}]$.

In summary, the results of this part of the analysis show that, in the case of "decreasing returns of scale", "many but bad" examples are associated with a smaller large-sample upper bound on the conditional generalization error than "few but good" ones. The opposite occurs for "increasing returns of scale", whereas the case of "constant returns of scale" is intermediate. These results are qualitatively in line with the ones obtained in Section 1 for the balanced case and in Gnecco and Nutarelli (2019b) and Gnecco and Nutarelli (2019c) for simpler linear regression problems, to which the ordinary/weighted least squares algorithms were applied. This depends on the fact that, in all these cases, the large-sample approximation of the conditional generalization error (or its large-sample upper bound) has the functional form $\frac{\sigma^2}{T}K_i$, where K_i is either a constant, or depends on optimization variables not related to both σ and T.

One can observe that, in order to discriminate among the three cases of the analysis reported above, it is not needed to know the exact values of the constants k and N, neither the expression of K_i as a function of the q_n. Moreover, to discriminate between the first two cases, it is not necessary to know the exact value of the positive constant α. Indeed, it suffices to know if α belongs, respectively, to the interval

$(0,1)$ or the interval $(1,+\infty)$. Finally, for this part of the analysis, knowledge of the probability distributions of the input examples associated with the different units is limited to the determination of the expressions of the constants $\lambda_{\min}(\boldsymbol{A}_n)$ involved in the optimization of the variables q_n.

Assuming that the constant terms $\lambda_{\min}(\boldsymbol{A}_n)$ and $\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}$ are known, optimal q_n° can be derived as follows. First, note that, for each fixed admissible choice of q_i, the optimization of the other q_n can be restated as follows:

$$\begin{aligned}\underset{\substack{q_n\in[q_{n,\min},q_{n,\max}],\\ n=1,\ldots,N,n\neq i}}{\text{maximize}} & \left(\lambda_{\min}(\boldsymbol{A}_i)q_i+\sum_{n=1,\ldots,N,n\neq i}\lambda_{\min}(\boldsymbol{A}_n)q_n\right)\\ \text{s.t.} & \sum_{n=1,\ldots,N,n\neq i} q_n=\bar{q}N-q_i\,.\end{aligned} \tag{1.58}$$

More precisely, an admissible choice for q_i is one for which $q_i\in[\hat{q}_{i,\min},\hat{q}_{i,\max}]$, where

$$\hat{q}_{i,\min}:=\max\{q_{i,\min},\bar{q}N-\sum_{n=1,\ldots,N,n\neq i} q_{n,\max}\}\,, \tag{1.59}$$

and

$$\hat{q}_{i,\max}:=\min\{q_{i,\max},\bar{q}N-\sum_{n=1,\ldots,N,n\neq i} q_{n,\min}\}\,. \tag{1.60}$$

The optimization problem (1.58) is a linear programming one, which can be reduced to a continuous knapsack problem (Martello and Toth, 1990, Section 2.2.1), after a rescaling of all its optimization variables and of their respective bounds. It is well known that, due to its particular structure, such a problem can be solved by the following greedy algorithm, which is divided into three steps (for simplicity of exposition, we assume that all the $\lambda_{\min}(\boldsymbol{A}_n)$ are different from each other):

1. first, the variables q_n are re-ordered according to decreasing values of the associated $\lambda_{\min}(\boldsymbol{A}_n)$. So, let $\breve{q}_n:=q_{\pi(n)}$ and $\breve{\boldsymbol{A}}_n:=\boldsymbol{A}_{\pi(n)}$, where the function $\pi:\{1,\ldots,N\}\to\{1,\ldots,N\}$ is a permutation satisfying $\lambda_{\min}(\breve{\boldsymbol{A}}_m)<\lambda_{\min}(\breve{\boldsymbol{A}}_n)$ for every $m\geq n$. Let also $\breve{i}=\pi(i)$;

2. starting from $\breve{q}_n=\breve{q}_{n,\min}$ for every $n\neq\breve{i}$, the first variable $\breve{q}_1$ (if $\breve{i}\neq 1$) is increased

until either the constraint $\sum_{n=1,\dots,N,n\neq\check{i}}\check{q}_n = \bar{q}N - \check{q}_{\check{i}}$, or the constraint $\check{q}_1 = \check{q}_{1,\max}$, is met; if $\check{i}=1$, then the procedure is applied to the second variable $\check{q}_2$;

3. step 2 is repeated for the successive variables (excluding $\check{q}_{\check{i}}$), stopping the first time the constraint $\sum_{n=1,\dots,N,n\neq\check{i}}\check{q}_n = \bar{q}N - \check{q}_{\check{i}}$ is met (this surely occurs, since q_i is admissible).

The resulting optimal q_n° (for $n=1,\dots,N$ with $n\neq i$) are parametrized by the remaining variable q_i. Then, the optimal value of the objective function of the optimization problem (1.58) is a real-valued function of q_i which, in the following, is denoted by $f_i(q_i)$. It follows from the procedure above that $f_i(q_i)$ is a continuous and piece-wise affine function of q_i, with piece-wise constant slopes $\lambda_{\min}(\check{\boldsymbol{A}}_{\check{i}}) - \lambda_{\min}(\check{\boldsymbol{A}}_{n(q_i)})$, where the choice of the index n is a function of q_i, and is such $\lambda_{\min}(\check{\boldsymbol{A}}_{n(q_i)})$ is a nonincreasing function of q_i. Hence, $f_i(q_i)$ is concave, and is nondecreasing for $q_i \leq \bar{q}N - \sum_{n=1}^{\check{i}-1}\check{q}_{n,\max}$, where $\check{q}_{n,\max} := q_{\pi(n),\max}$, and nonincreasing otherwise.
Exploiting the results above, the optimal value of q_i for the original optimization problem (1.57) is obtained by solving the following optimization problem:

$$\underset{q_i\in[\hat{q}_{i,\min},\hat{q}_{i,\max}]}{\text{minimize}} \left(\frac{1}{q_i} + \frac{\mathbb{E}\left\{ \left\| \left(\mathbb{E}\left\{ \boldsymbol{x}_{i,1} \right\} - \boldsymbol{x}_i^{test} \right) \right\|_2^2 \right\}}{f_i(q_i)} \right). \tag{1.61}$$

This is a convex optimization problem, since the function $\frac{1}{q_i}$ is convex, whereas the function

$$\frac{\mathbb{E}\left\{ \left\| \left(\mathbb{E}\left\{ \boldsymbol{x}_{i,1} \right\} - \boldsymbol{x}_i^{test} \right) \right\|_2^2 \right\}}{f_i(q_i)} \tag{1.62}$$

is of the form $h(f_i)$, where f_i is concave and h is convex and nonincreasing, so $h(f_i)$ is convex Vandenberghe and Boyd (2004, Section 3.2). After solving the optimization problem (1.61), the optimal values of the other q_n for the original optimization problem (1.57) are obtained as a consequence of the three steps detailed above.

It follows from the reasoning above that the structure of the optimal solutions q_n° is as follows. First, there exists a threshold $\bar{\lambda}^\circ > 0$ such that

i) for any $n\neq i$ with $\lambda_{\min}(\boldsymbol{A}_n) > \bar{\lambda}^\circ$, q_n° is equal to its maximum admissible value $q_{n,\max}$;

ii) for any $n \neq i$ with $\lambda_{\min}(\boldsymbol{A}_n) < \bar{\lambda}^\circ$, q_n° is equal to its minimum admissible value $q_{n,\min}$;

iii) for at most one unit $n \neq i$ (for which $\lambda_{\min}(\boldsymbol{A}_n) = \bar{\lambda}^\circ$, provided that there exists one value of n for which this condition holds), q_n° belongs to the interior of the interval $[q_{n,\min}, q_{n,\max}]$.

Moreover,

iv) if $\left(\bar{q}N - \sum_{n=1}^{\check{i}-1} \check{q}_{n,\max}\right) \geq \hat{q}_{i,\max}$, then

$$q_i^\circ = \hat{q}_{i,\max}\,, \tag{1.63}$$

and

$$\bar{\lambda}^\circ \in (0, \lambda_{\min}(\boldsymbol{A}_i))\,; \tag{1.64}$$

v) if $\left(\bar{q}N - \sum_{n=1}^{\check{i}-1} \check{q}_{n,\max}\right) < \hat{q}_{i,\max}$, then

$$q_i^\circ \in \left[\left(\bar{q}N - \sum_{n=1}^{\check{i}-1} \check{q}_{n,\max}\right), \hat{q}_{i,\max}\right], \tag{1.65}$$

and

$$\bar{\lambda}^\circ > \lambda_{\min}(\boldsymbol{A}_i)\,. \tag{1.66}$$

Finally, it is worth observing that the structure highlighted above for the optimal solutions q_n° and c° (the latter reported under Equation (1.57)), which is valid for any fixed value of $\bar{q}$, can be useful to solve the modification of the optimization problem (1.57) obtained in case the constraint (1.52) is replaced by

$$\bar{q}_{\min}N \leq \sum_{n=1}^{N} q_n \leq \bar{q}_{\max}N\,, \tag{1.67}$$

for some given $\bar{q}_{\min}, \bar{q}_{\max} \in (0,1]$, with $\bar{q}_{\min} < \bar{q}_{\max}$.

2.4 Discussion

In this Section we introduced a baseline scenario displaying the form of the generalization error for the baseline case of balanced fixed effects in Section 1. Then, in Section 2, the optimal trade-off between sample size, precision of supervision, and selection probabilities, has been studied with specific reference to a quite general linear model of input-output relationship representing unobserved heterogeneity in the data, namely the unbalanced fixed effects panel data model. First, we have analyzed its conditional generalization error, then we have minimized a large-sample upper bound on it with respect to some of its parameters. We have proved that, under suitable assumptions, "many but bad" examples provide a smaller upper bound on the conditional generalization error than "few but good" ones, whereas in other cases the opposite occurs. The choice between "many but bad" and "few but good" examples plays an important role, when better supervision implies higher costs. We also show that the problem of minimizing the generalization error provides similar results in Sections 1 and 2.

The next Section will extend the obtained results to the case of correlated errors. The theoretical results obtained in this Section could be applied to the acquisition design of unbalanced panel data related to several fields, such as biostatistics, econometrics, educational research, engineering, neuroscience, political science, and sociology. Moreover, the analysis of the large-sample case could be extended to deal with large N, or with both large N and T. These cases would be of interest for their potential applications in microeconometrics (Cameron and Trivedi, 2005). Another possible extension concerns the introduction, in the noise model, of a subset of not controllable parameters (beyond the controllable one, i.e., the noise variance), which could be estimated from a subset of training data. As a final extension, one could investigate and optimize the trade-off between sample size and precision of supervision (and possibly, also selection probabilities) for the random effects panel data model (Greene, 2003, Section 13). This is also commonly applied in the analysis of economic data, and differs from the fixed effects panel data model in that its parameters are considered as random variables. In the present context, however, a possible advantage of the fixed effects panel data model is that it also allows one to obtain estimates of the individual constants η_n (see Equation (1.31)), which appear in the expression (1.32) of

the conditional generalization error. Moreover, the application of the random effects model to the unbalanced case requires stronger assumptions than the one of the fixed effects model (Wooldridge, 2010, Section 17).

Chapter 2

Optimal data collection design in machine learning: the case of correlated measurement errors[1]

Introduction

In Chapter 1 we investigated the optimal trade-off between training set size and precision of supervision in the general linear context of the input/output relationship, namely, the baseline fixed effects panel data model.

In order to increase the applicability of the analysis carried out in the previous Chapter, in this Chapter we extend it thoroughly in the following directions. First, in Chapter 1 we investigated only the case in which the measurements errors of observations associated with the same unit are mutually independent. In this paper, we extend such analysis to the case of dependent measurement errors. Moreover, differently from Chapter 1, we confirm the validity of the obtained theoretical results numerically. Further, in Chapter 1, the optimal trade-off between training set size

[1]This chapter is partially based on Gnecco, G., Nutarelli, F. & Selvi, D. Optimal data collection design in machine learning: the case of the fixed effects generalized least squares panel data model. Mach Learn 110, 1549–1584 (2021). https://doi.org/10.1007/s10994-021-05976-x

and precision of supervision was analyzed only for a fixed number of units, assuming that the number of observations associated with the same unit is large enough to justify a large-sample approximation with respect to the number of observations. In the last part of this Chapter, we consider additionally the cases of a large-sample approximation with respect to the number of units, and of a large-sample approximation with respect to both the number of units and the number of observations per unit.

It is important to highlight, that the model described in this Chapter does work for balanced data differently from Chapter 1. In summary, Section 2 of Chapter 1 extends the baseline model in that it takes unbalanced data while the baseline considers only balanced data, while Chapter 2 innovates the baseline model in that it considers the case of correlated measurement errors (while the baseline does not).

In line with the results of the theoretical analyses made in Chapter 1 and Gnecco and Nutarelli (2019b) and Gnecco and Nutarelli (2019c) for simpler linear regression models, we show that, also for the more applicable fixed effects generalized least squares panel data model, the following holds in general: when the precision of each supervision (i.e., the reciprocal of the conditional variance of the output variable, given the associated unit and input variables) increases less than proportionally versus an increase of the supervision cost per training example, the minimum (large-sample approximation of the) generalization error (conditioned on the training input data) is obtained in correspondence of the smallest supervision cost per example (hence, of the largest number of examples); when that precision increases more than proportionally versus an increase of the supervision cost per example, the optimal supervision cost per example is the largest one (which corresponds to the smallest number of examples).

As a further novelty, in the present Chapter, the number of training examples can be varied either by increasing the number of observations per unit, or the number of units, or both.

The Chapter is structured as follows. Section 2 provides a background on the fixed effects generalized least squares panel data model. Section 2 presents the analysis of its conditional generalization error, and of the large-sample approximation of the latter with respect to time. Section 2 formulates and solves an optimization problem

we propose in order to provide an optimal trade-off between training set size and precision of supervision for the fixed effects generalized least squares panel data model, using the large-sample approximation above. Section 4 presents some numerical results, which validate the theoretical ones. Finally, Section 2 discusses some possible applications and extensions of the theoretical results obtained in the work. Some technical proofs and remarks about the extension of the analysis made in the paper to other large-sample settings are reported in the Appendix 2.

Background

In this Section, we recall some basic facts about the following Fixed Effects Generalized Least Squares (FEGLS) panel data model (see, e.g., Wooldridge, 2010, Chapter 10). Specifically, we refer to the baseline model of Eq.(1.1) which we report again below:

$$y_{n,t} := \eta_n + \underline{\beta}' \underline{x}_{n,t}, \text{ for } n = 1, \ldots, N, t = 1, \ldots, T, \tag{2.1}$$

As in the baseline model, the outputs $y_{n,t}$ are actually unavailable. We change here the notation of their noisy measuraments to conform to FEGLS literature. In particular, $\tilde{y}_{n,t}$ are now denoted as $z_{n,t}$. The additive noise model, therefore, turns out to be:

$$z_{n,t} := y_{n,t} + \varepsilon_{n,t}, \text{ for } n = 1, \ldots, N, t = 1, \ldots, T, \tag{2.2}$$

where, for any n, the $\varepsilon_{n,t}$ are identically distributed and possibly dependent random variables, having mean 0, and are further independent from all the $\underline{x}_{n,t}$. For any two units $n \neq m$ and any two time instants $t_1, t_2 \in \{1, \ldots, T\}$, ε_{n,t_1} and ε_{m,t_2} are assumed to be independent. Hence, only the possibility of temporal dependence for the measurement errors associated with the same unit is considered in the following, in line with several works in the literature (see, e.g., Bhargava, Franzini, and Narendranathan, 1982 and Wooldridge, 2010, Section 10.5.5).

For each unit n, let $X_n \in \mathbb{R}^{T \times p}$ be the matrix whose rows are the transposes of the $\underline{x}_{n,t}$; further, let $\underline{z}_n \in \mathbb{R}^T$ be the column vector which collects the noisy measurements $z_{n,t}$, and $\underline{\varepsilon}_n \in \mathbb{R}^T$ the column vector which collects the measurement noises $\varepsilon_{n,t}$. The input/corrupted output pairs $(\underline{x}_{n,t}, \underline{z}_{n,t})$, for $n = 1, \ldots, N$, $t = 1, \ldots, T$, are used to train the FEGLS model, i.e., to estimate its parameters.

The following first-order serial covariance form is assumed (see, e.g., Bhargava,

Franzini, and Narendranathan, 1982 and Wooldridge, 2010, Section 10.5.5) for the (unconditional) covariance matrix of the vector of measurement noises associated with the n-th unit[2], where $\sigma > 0$ and $\rho \in (-1, 1)$ hold (here, $\mathbb{E}$ denotes the expectation operator):

$$\Lambda := \sigma^2 \Psi := \mathrm{Var}\left(\underline{\varepsilon}_n\right) = \mathbb{E}\{\underline{\varepsilon}_n \underline{\varepsilon}_n'\} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho^2 & \rho & 1 \end{bmatrix} \in \mathbb{R}^{T \times T}, \tag{2.3}$$

which is a symmetric and positive-definite matrix. In other words, the measurement noise is assumed to be generated by a first-order autoregressive ($AR(1)$) process (Ruud et al., 2000, Section 25.2). In the particular case of uncorrelated ($\rho = 0$) and independent measurement noises, one obtains the model considered in Chapter 1.

Let the matrix $Q_T \in \mathbb{R}^{T \times T}$ be defined as

$$Q_T := I_T - \frac{1}{T} \underline{1}_T \underline{1}_T', \tag{2.4}$$

where $I_T \in \mathbb{R}^{T \times T}$ is the identity matrix, and $\underline{1}_T \in \mathbb{R}^T$ a column vector whose elements are all equal to 1. One can check that Q_T is a symmetric and idempotent matrix (i.e., $Q_T' = Q_T = Q_T^2$), and its eigenvalues are 0 with multiplicity 1, and 1 with multiplicity $T - 1$. Hence, for each unit n, one can define, similarly to Chapter 1,

$$\ddot{X}_n := Q_T X_n = \begin{bmatrix} \underline{x}_{n,1} - \frac{1}{T} \sum_{t=1}^T \underline{x}_{n,t} \\ \underline{x}_{n,2} - \frac{1}{T} \sum_{t=1}^T \underline{x}_{n,t} \\ \cdots \\ \underline{x}_{n,T} - \frac{1}{T} \sum_{t=1}^T \underline{x}_{n,t} \end{bmatrix}, \tag{2.5}$$

$$\ddot{\underline{z}}_n := Q_T \underline{z}_n = \begin{bmatrix} z_{n,1} - \frac{1}{T} \sum_{t=1}^T z_{n,t} \\ z_{n,2} - \frac{1}{T} \sum_{t=1}^T z_{n,t} \\ \cdots \\ z_{n,T} - \frac{1}{T} \sum_{t=1}^T z_{n,t} \end{bmatrix}, \tag{2.6}$$

[2]An important implication of first-order serial covariance in noise terms is the unreliability of classical test statistics, based on the assumption of uncorrelated noises (see, e.g., Im et al., 1999). To deal with this issue, the usual approach adopted in the literature consists in explicitly taking into account the form (2.3) for the covariance matrix of the zero-mean vector of measurement noises associated with each unit.

and

$$\ddot{\underline{\varepsilon}}_n := Q_T \underline{\varepsilon}_n = \begin{bmatrix} \varepsilon_{n,1} - \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{n,t} \\ \varepsilon_{n,2} - \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{n,t} \\ \cdots \\ \varepsilon_{n,T} - \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{n,t} \end{bmatrix}, \tag{2.7}$$

which represent, respectively, the matrix of time de-meaned training inputs, the vector of time de-meaned corrupted training outputs, and the vector of time de-meaned measurements noises. The goal of time de-meaning is to obtain a derived dataset where the fixed effects are removed, making it possible to estimate first the vector $\underline{\beta}$, then - turning back to the original dataset - the fixed effects η_n. The covariance matrix $\mathbb{E}\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}_n'\}$ has the expression

$$\Omega := \sigma^2 \Phi := \mathrm{Var}\,(\ddot{\underline{\varepsilon}}_n) = \mathbb{E}\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}_n'\} = Q_T \mathbb{E}\{\underline{\varepsilon}_n \underline{\varepsilon}_n'\} Q_T' = Q_T \Lambda Q_T' = \sigma^2 Q_T \Psi Q_T', \tag{2.8}$$

which is symmetric and positive semi-definite, and has rank $T-1 < T$ (Wooldridge, 2010). Although this deficient rank prevents the application of the most usual approach to Generalized Least Squares (GLS) estimation, based on the inversion of the covariance matrix Ω (which in this case cannot be inverted), one can still apply GLS by projecting Eqs. (2.1) and (2.2) onto the orthogonal complement L of the vector $\underline{1}_T$ by using Q_T, then solving a standard GLS problem on L (Aitken, 1936). This is formally obtained by replacing the inverse of the covariance matrix with its Moore-Penrose pseudoinverse[3] (denoted by Ω^+), as made in the context of FEGLS estimation in Kiefer (1980) and Im et al. (1999). More precisely, assuming the invertibility of the matrix $\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{X}_n$ (see Remark 6 for a justification of this assumption), the

[3] It is recalled here from Strang (1993) that the Moore-Penrose pseudoinverse M^+ of a matrix $M \in \mathbb{R}^{T \times T}$ inverts a special restriction of the linear application represented by the matrix M, whose domain and codomain are restricted, respectively, to the row space of M and to the column space of M (which coincide in the case of a symmetric matrix). For a matrix $M \in \mathbb{R}^{T \times T}$ with singular value decomposition $M = U \Sigma V'$ (where $U, V \in \mathbb{R}^{T \times T}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{T \times T}$ is a diagonal matrix whose non-zero entries are the singular values of M), the singular value decomposition of its Moore-Penrose pseudoinverse is $M^+ = U \Sigma^+ V'$ (where $\Sigma^+ \in \mathbb{R}^{T \times T}$ is a diagonal matrix whose non-zero entries are the reciprocals of the singular values of M). Finally, in the particular case in which M is symmetric and positive semi-definite (e.g., when it is a covariance matrix), its singular values coincide with its positive eigenvalues, $U \in \mathbb{R}^{T \times T}$ is an orthogonal matrix whose columns are its eigenvectors, $V' = U^{-1}$, and M^+ is symmetric and positive semi-definite. The concept of Moore-Penrose pseudoinversion can be extended to the cases of rectangular matrices and matrices with complex entries, but such extensions are not needed in this work.

FEGLS estimate of $\underline{\beta}$ is

$$\underline{\hat{\beta}}_{FEGLS} = \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{\underline{z}}_n\right). \tag{2.9}$$

The estimate $\underline{\hat{\beta}}_{FEGLS}$ in (2.9) can be interpreted as the GLS estimate of $\underline{\beta}$ obtained by replacing the original input/corrupted output training data with their de-meaned versions reported above. It is worth observing that the training input/corrupted output pairs $\left(\underline{x}_{n,t}, z_{n,t}\right)$ $(n = 1, \ldots, N, t = 1, \ldots, T)$ are all used to estimate $\underline{\beta}$.

Remark 5. *Another commonly used approach to deal with the issue above is to drop one of the time periods from the analysis, in order to get an invertible covariance matrix. It can be rigorously proved (see, e.g. Im et al., 1999, Theorem 4.3) that this second approach is equivalent to the one based on the Moore-Penrose pseudoinverse (producing exactly the same FEGLS estimate), and that it does not matter which time period is dropped, as the resulting GLS estimator has always the same form. Therefore, dropping the last row of Q_T, one gets the matrix $\tilde{Q}_T \in \mathbb{R}^{(T-1)\times T}$, from which one obtains the matrix $\tilde{\ddot{X}}_n := \tilde{Q}_T X_n \in \mathbb{R}^{(T-1)\times p}$, the column vector $\tilde{\ddot{\underline{z}}}_n := \tilde{Q}_T \underline{z}_n \in \mathbb{R}^{T-1}$, and the column vector $\tilde{\ddot{\underline{\varepsilon}}}_n := \tilde{Q}_T \underline{\varepsilon}_n \in \mathbb{R}^{T-1}$. Moreover, denoting by $\tilde{X}_n \in \mathbb{R}^{(T-1)\times p}$, $\tilde{\underline{z}}_n \in \mathbb{R}^{T-1}$, and $\tilde{\underline{\varepsilon}}_n \in \mathbb{R}^{T-1}$ the matrix and the vectors obtained by removing the last row, respectively, from X_n, $\underline{z}_n$, and $\underline{\varepsilon}_n$, one gets*

$$\tilde{\Omega} := \mathbb{E}\{\tilde{\ddot{\underline{\varepsilon}}}_n \tilde{\ddot{\underline{\varepsilon}}}_n'\} = \tilde{Q}_T \mathbb{E}\{\underline{\varepsilon}_n \underline{\varepsilon}_n'\} \tilde{Q}_T' = \tilde{Q}_T \Lambda \tilde{Q}_T', \tag{2.10}$$

which is, differently from Ω, an invertible matrix, with inverse $\tilde{\Omega}^{-1} = (\tilde{Q}_T \Lambda \tilde{Q}_T')^{-1}$. The resulting FEGLS estimate is

$$\underline{\hat{\beta}}^{alt}_{FEGLS} = \left(\sum_{n=1}^{N} \tilde{\ddot{X}}_n' \tilde{\Omega}^{-1} \tilde{\ddot{X}}_n\right)^{-1} \left(\sum_{n=1}^{N} \tilde{\ddot{X}}_n' \tilde{\Omega}^{-1} \tilde{\ddot{\underline{z}}}_n\right). \tag{2.11}$$

(see, e.g., Wooldridge, 2010). The FEGLS estimate $\underline{\hat{\beta}}_{FEGLS}$ and the alternative one $\underline{\hat{\beta}}^{alt}_{FEGLS}$ are actually identical (Im et al., 1999, Theorem 4.3). This equivalence is obtained by expressing such estimates in terms of the original variables before de-meaning, then exploiting the proof of Im et al. (1999, Theorem 4.3), which shows that $Q_T' \Omega^+ Q_T = \tilde{Q}_T' \tilde{\Omega}^{-1} \tilde{Q}_T$ (this still holds if an observation different from the last one is dropped, and $\tilde{Q}_T$ is redefined accordingly).

The FEGLS estimates of the η_n (also called fixed effects residuals, see Wooldridge

(2010)) are

$$\hat{\eta}_{n,FEGLS} := \frac{1}{T}\sum_{t=1}^{T}\left(z_{n,t} - \hat{\underline{\beta}}'_{FEGLS}\underline{x}_{n,t}\right). \tag{2.12}$$

They are obtained by subtracting the estimate $\hat{\underline{\beta}}'_{FEGLS}\underline{x}_{n,t}$ of $\underline{\beta}'\underline{x}_{n,t}$ from each corrupted output $z_{n,t}$, then performing an empirical average, limiting to training data associated with the unit n. The FEGLS estimates reported in Eq. (2.12) are motivated by the fact that the η_n are constants, whereas the $\varepsilon_{n,t}$ have mean 0.

By taking expectations, it readily follows from their definitions that the estimates (2.9) and (2.12) are conditionally unbiased with respect to the training input data $\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}$, i.e., that

$$\mathbb{E}\left\{\left(\hat{\underline{\beta}}_{FEGLS} - \underline{\beta}\right)|\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} = \underline{0}_p\,, \tag{2.13}$$

where $\underline{0}_p \in \mathbb{R}^p$ is a column vector whose elements are all equal to 0, and, for any $i = 1,\ldots,N$,

$$\mathbb{E}\left\{(\hat{\eta}_{i,FEGLS} - \eta_i)\,|\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} = 0\,. \tag{2.14}$$

Finally, the covariance matrix of $\hat{\underline{\beta}}_{FEGLS}$, conditioned on the training input data, is

$$\mathrm{Var}\left(\hat{\underline{\beta}}_{FEGLS}|\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right) \;=\; \left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\ddot{X}_n\right)^{-1}. \tag{2.15}$$

Conditional generalization error and its large-sample approximation in the case of correlated errors

The goal of this Section is to analyze the generalization error associated with the FEGLS estimates (2.9) and (2.12), conditioned on the training input data, by providing its large-sample approximation. Then, in Section 2, the resulting expression is optimized, after choosing suitable models for the standard deviation σ of the measurement noise and for the time horizon, which is chosen in such a way it satisfies a suitable budget constraint.

First, we express the generalization error or expected risk for the i-th unit ($i = 1,\ldots,N$), conditioned on the training input data, by

$$R_i\left(\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right) := \mathbb{E}\left\{\left(\hat{\eta}_{i,FEGLS} + \hat{\underline{\beta}}'_{FEGLS}\underline{x}_i^{test} - \eta_i - \underline{\beta}'\underline{x}_i^{test}\right)^2 \Big|\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\}, \tag{2.16}$$

where $\underline{x}_i^{test} \in \mathbb{R}^p$ is independent from the training data. It is the expected mean

squared error of the prediction of the output associated with a test input, conditioned on the training input data.
As shown in Appendix 2, we can express the conditional generalization error (2.16) as follows, highlighting its dependence on σ^2:

$$
\begin{aligned}
& R_i\left(\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right) \\
= \; & \frac{\sigma^2}{T^2}\underline{1}_T' X_i \left(\sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} X_i' \underline{1}_T + \frac{\sigma^2}{T^2}\underline{1}_T' \Psi \underline{1}_T \\
& - \frac{2\sigma^2}{T^2}\underline{1}_T' X_i \left(\sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \ddot{X}_i' \Phi^+ Q_T \Psi \underline{1}_T + \sigma^2 \mathbb{E}\left\{ \left(\underline{x}_i^{test}\right)' \left(\sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \underline{x}_i^{test} | \{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T} \right\} \\
& - \frac{2\sigma^2}{T}\underline{1}_T' X_i \left(\sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \mathbb{E}\left\{\underline{x}_i^{test}\right\} + \frac{2\sigma^2}{T} (Q_T \Psi \underline{1}_T)' \Phi^+ \ddot{X}_i \left(\sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \mathbb{E}\left\{\underline{x}_i^{test}\right\}, \quad (2.17)
\end{aligned}
$$

where some computations (reported in Appendix 2) show that

$$
\underline{1}_T' \Psi \underline{1}_T = T + 2T\left(\frac{1-\rho^T}{1-\rho} - 1\right) - \frac{2\rho}{1-\rho}\left(-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1\right), \quad (2.18)
$$

and

$$
\begin{aligned}
\underline{u}_T \; := \; & Q_T \Psi \underline{1}_T \\
= \; & \begin{bmatrix}
\cancel{1} + \rho + \rho^2 + \rho^3 + \rho^4 + \cdots & +\rho^{T-1} & \cancel{-1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1]}{T(1-\rho)} \\
\rho + \cancel{1} + \rho + \rho^2 + \rho^3 + \cdots & +\rho^{T-2} & \cancel{-1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1]}{T(1-\rho)} \\
\rho^2 + \rho + \cancel{1} + \rho + \rho^2 + \cdots & +\rho^{T-3} & \cancel{-1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1]}{T(1-\rho)} \\
\cdots & \cdots & \cdots \\
\rho^{T-3} + \rho^{T-4} + \cdots + \rho + \cancel{1} + \rho & +\rho^2 & \cancel{-1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1]}{T(1-\rho)} \\
\rho^{T-2} + \rho^{T-3} + \cdots + \rho^2 + \rho + \cancel{1} & +\rho & \cancel{-1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1]}{T(1-\rho)} \\
\rho^{T-1} + \rho^{T-2} + \cdots + \rho^3 + \rho^2 + \rho & +\cancel{1} & \cancel{-1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1]}{T(1-\rho)}
\end{bmatrix}. \quad (2.19)
\end{aligned}
$$

Next, we obtain a large-sample approximation of the conditional generalization error (2.17) with respect to T, for a fixed number of units N. Such an approximation is useful, e.g., in the application of the model to macroeconomics data, for which it is common to investigate the case of a large horizon T.

Under mild conditions (e.g., if for the unit i the $\underline{x}_{i,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4), the following conver-

gences in probability[4] hold (their proofs are reported in Appendix 2):

$$\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\underline{1}_T'X_i=\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)', \tag{2.20}$$

$$\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\ddot{X}_i'\Phi^+Q_T\Psi\underline{1}_T=\underline{0}_p\,. \tag{2.21}$$

Similarly, if for each fixed unit n the $\underline{x}_{n,t}$ are mutually independent, identically distributed[5], and have finite moments up to the order 4, and one makes the additional assumption (whose validity is discussed extensively in Appendix 2) that

$$\lim_{T\to\infty}\|\Phi^+-Q_T\Psi^{-1}Q_T'\|_2=0 \tag{2.22}$$

(where, for a symmetric matrix $M\in\mathbb{R}^{T\times T}$, $\|M\|_2=\max\limits_{t=1,\ldots,T}|\lambda_t(M)|$ denotes its spectral norm), then also the following convergence in probability holds:

$$\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{n=1}^{N}\ddot{X}_n'\Phi^+\ddot{X}_n=A_N\,, \tag{2.23}$$

where

$$A_N=A_N':=\frac{1+\rho^2}{1-\rho^2}\sum_{n=1}^{N}\mathbb{E}\left\{\left(\underline{x}_{n,1}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)\left(\underline{x}_{n,1}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)'\right\} \tag{2.24}$$

is a symmetric and positive semi-definite matrix. In the following, its positive definiteness (hence, its invertibility) is also assumed.

Remark 6. *The existence of the probability limit (2.23) and the assumed positive definiteness of the matrix A_N guarantee that the invertibility of the matrix $\sum_{n=1}^{N}\ddot{X}_n'\Phi^+\ddot{X}_n$ (see the invertibility assumption before Eq. (2.9)) holds with probability near* 1 *for large T.*

When (2.20), (2.21), and (2.23) hold, inserting such probability limits in Eq. (2.17), one gets the following large-sample approximation of the conditional generalization error (2.17) with respect to T:

$$\begin{aligned}(2.17)\ \simeq\ & \frac{\sigma^2}{T}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)'A_N^{-1}\mathbb{E}\left\{\underline{x}_{i,1}\right\}+\frac{\sigma^2}{T}\frac{1+\rho}{1-\rho}\\ &+\frac{\sigma^2}{T}\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)'A_N^{-1}\underline{x}_i^{test}\right\}-2\frac{\sigma^2}{T}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)'A_N^{-1}\mathbb{E}\left\{\underline{x}_i^{test}\right\}\end{aligned}$$

[4]We recall that a sequence of random real matrices M_T, $T=1,2,\ldots$, converges in probability to the real matrix M if, for every $\varepsilon>0$, $\mathrm{Prob}\left(\|M_T-M\|>\varepsilon\right)$ (where $\|\cdot\|$ is an arbitrary matrix norm) tends to 0 as T tends to $+\infty$. In this case, we write $\operatorname*{plim}\limits_{T\to+\infty}M_T=M$.

[5]This does not exclude the possibility for the $\underline{x}_{n,t}$ and $\underline{x}_{m,t}$ associated with different units n and m to be dependent/not identically distributed.

$$= \frac{\sigma^2}{T}\left(\frac{1+\rho}{1-\rho}+\mathbb{E}\left\{\left\|A_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}-\underline{x}_i^{test}\right)\right\|_2^2\right\}\right), \tag{2.25}$$

where, for a vector $\underline{v} \in \mathbb{R}^p$, $\|\underline{v}\|_2$ denotes its l_2 (Euclidean) norm, and $A_N^{-\frac{1}{2}}$ is the principal square root (i.e., the symmetric and positive definite square root) of the symmetric and positive definite matrix A_N^{-1}. Eq. (2.25) is obtained taking into account that, as a consequence of the Continuous Mapping Theorem (Florescu, 2014, Theorem 7.33), the probability limit of the product of two random variables equals the product of their probability limits, when the latter two exist. By doing this, the third and sixth terms of Eq. (2.17) cancel out due to Eq. (2.21), whereas the second term is computed using Eq. (2.18).

Interestingly, the large-sample approximation (2.25) has the form $\frac{\sigma^2}{T}K_i$, where

$$K_i := \left(\frac{1+\rho}{1-\rho}+\mathbb{E}\left\{\left\|A_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}-\underline{x}_i^{test}\right)\right\|_2^2\right\}\right) \tag{2.26}$$

is a positive constant (possibly, a different constant for each unit i). This simplifies the analysis of the trade-off between training set size and precision of supervision performed in the next Section, since one does not need to compute the exact expression of K_i to find the optimal trade-off.

In Appendix 2, an extension of the analysis made above is presented, by considering, respectively, the case of large N, and the one in which both N and T are large.

Optimal trade-off between training set size and precision of supervision for the fixed effects generalized least squares panel data model under the large-sample approximation

In this Section, we are interested in optimizing the large-sample approximation (2.25) of the conditional generalization error when the variance σ^2 is modeled as a decreasing function of the supervision cost per example c, and there is an upper bound C on the total supervision cost NTc associated with the whole training set. In the analysis, N is fixed, and T is chosen as $\left\lfloor \frac{C}{Nc} \right\rfloor$. Moreover, the supervision cost per example c is allowed to take values on the interval $[c_{\min}, c_{\max}]$, where $0 < c_{\min} < c_{\max}$, so that the resulting T belongs to $\left\{\left\lfloor \frac{C}{Nc_{\max}} \right\rfloor, \ldots, \left\lfloor \frac{C}{Nc_{\min}} \right\rfloor\right\}$. In the following, C is supposed to be sufficiently large, so that the large-sample approximation (2.25) can be assumed to hold for every $c \in [c_{\min}, c_{\max}]$.

Consistently with Chapter 1 and Gnecco and Nutarelli (2019b) and Gnecco and

Nutarelli (2019c), we adopt the following model for the variance σ^2, as a function of the supervision cost per example c:

$$\sigma^2(c) = kc^{-\alpha}, \tag{2.27}$$

where $k, \alpha > 0$. For $0 < \alpha < 1$, the precision of each supervision is characterized by "decreasing returns of scale" with respect to its cost because, if one doubles the supervision cost per example c, then the precision $1/\sigma^2(c)$ becomes less than two times its initial value (or equivalently, the variance $\sigma^2(c)$ becomes more than one half its initial value). Conversely, for $\alpha > 1$, there are "increasing returns of scale" because, if one doubles the supervision cost per example c, then the precision $1/\sigma^2(c)$ becomes more than two times its initial value (or equivalently, the variance $\sigma^2(c)$ becomes less than one half its initial value). The case $\alpha = 1$ is intermediate and refers to "constant returns of scale". In all the cases above, the precision of each supervision increases by increasing the supervision cost per example c. Finally, it is worth observing that, according to the model (2.3) for the covariance matrix of the vector of measurement noises, the correlation coefficient between successive measurement noises does not depend on c.

Concluding, under the assumptions above, the optimal trade-off between the training set size and the precision of supervision for the fixed effects generalized least squares panel data model is modeled by the following optimization problem:

$$\underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}} \; K_i k \frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor - 1}. \tag{2.28}$$

By a similar argument as in the proof of the baseline model in Chapter 1 (see Gnecco and Nutarelli, 2019c, Proposition 3.2 for more details), when C is sufficiently large, the objective function $CK_i k \frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor - 1}$ of the optimization problem (2.28), rescaled by the multiplicative factor C, can be approximated, with a negligible error in the maximum norm on $[c_{\min}, c_{\max}]$, by $NK_i k c^{1-\alpha}$. In order to illustrate this issue, Figure 2 shows the behavior of the rescaled objective functions $CK_i k \frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor - 1}$ and $NK_i k c^{1-\alpha}$ for the three cases $0 < \alpha = 0.5 < 1$, $\alpha = 1.5 > 1$, and $\alpha = 1$ (the values of the other parameters are $k = 0.5$, $K_i = 2$, $N = 10$, $C = 200$, $c_{\min} = 0.4$, and $c_{\max} = 0.8$). The additional approximation $CNK_i k \frac{c^{1-\alpha}}{C - Nc}$ (which differs negligibly from $NK_i k c^{1-\alpha}$ for large C) is also reported in the figure.

Concluding, under the approximation above, one can replace the optimization prob-

lem (2.28) with

$$\underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}} \, N K_i k c^{1-\alpha}, \tag{2.29}$$

whose optimal solutions c° have the following expressions:

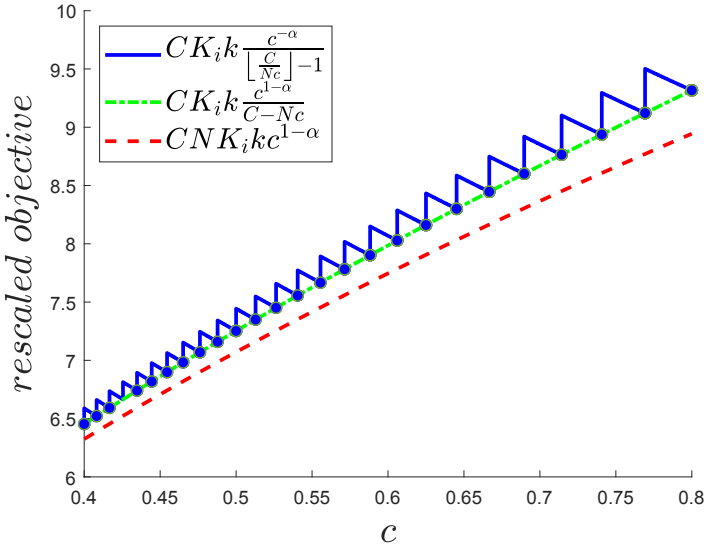
i) if $0 < \alpha < 1$ ("decreasing returns of scale"): $c^\circ = c_{\min}$;

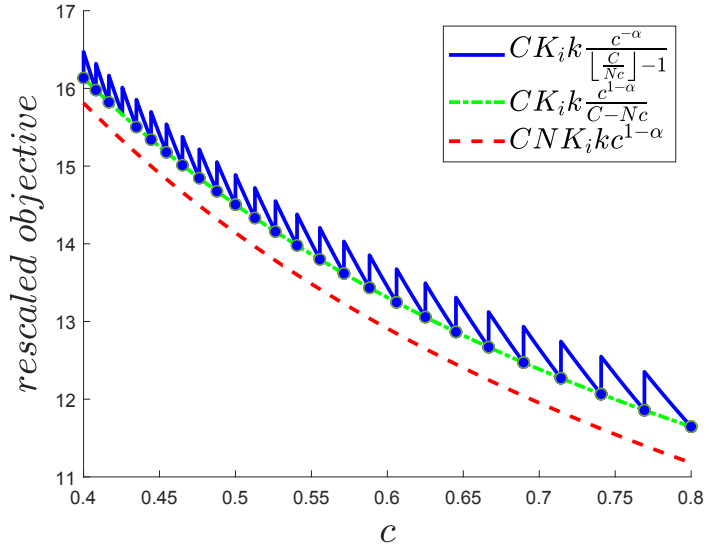
ii) if $\alpha > 1$ ("increasing returns of scale"): $c^\circ = c_{\max}$;

iii) if $\alpha = 1$ ("constant returns of scale"): $c^\circ =$ any cost c in the interval $[c_{\min}, c_{\max}]$.

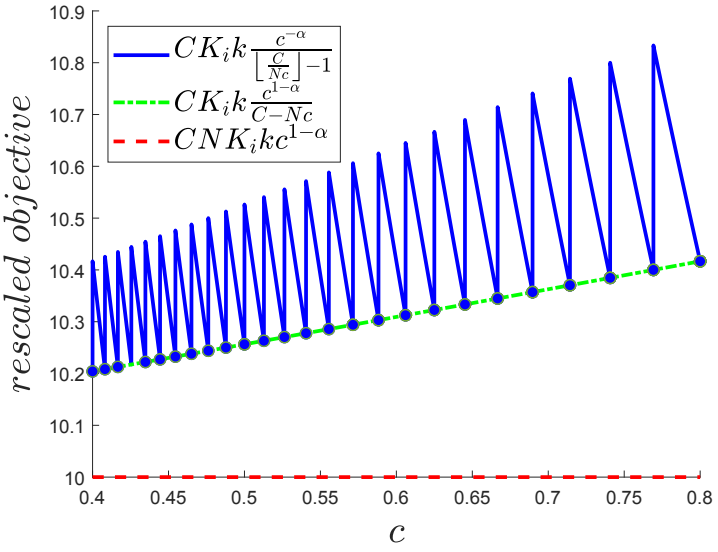
In summary, the results of the analysis show that, in the case of "decreasing returns of scale", "many but bad" examples are associated with a smaller conditional generalization error than "few but good" ones. The opposite occurs for "increasing returns of scale", whereas the case of "constant returns of scale" is intermediate. These results are qualitatively in line with the ones obtained in Chapter 1 and in Gnecco and Nutarelli (2019b) and Gnecco and Nutarelli (2019c) for simpler linear regression problems, to which different regression algorithms were applied (respectively, ordinary least squares, weighted least squares, and fixed effects ordinary least squares). This depends on the fact that, in all these cases, the large-sample approximation of the conditional generalization error has the same functional form $\frac{\sigma^2}{T} K_i$ (although different positive constants K_i are involved in the various cases).

We remember here, as mentioned in Chapter 1, that, in order to discriminate among the three cases of the analysis reported above, one does not need to know the exact values of the constants ρ, k, K_i, and N. Moreover, to discriminate between the first two cases, it is not necessary to know the exact value of the positive constant α (indeed, it suffices to know if α belongs, respectively, to the interval $(0, 1)$ or the one $(1, +\infty)$). Interestingly, no precise knowledge of the probability distributions of the input examples (one for each unit) is needed. In particular, different probability distributions may be associated with different units, without affecting the results of the analysis. Finally, the same conclusions as above are reached if the objective function in (2.29) is replaced by the summation of the large-sample approximation of the conditional generalization error over all the N units. In that case, the constant K_i in (2.29) is replaced by $K := \sum_{i=1}^{N} K_i$.


$CK_ik\frac{c^{-\alpha}}{\left\lfloor\frac{C}{Nc}\right\rfloor-1}$
$CK_ik\frac{c^{1-\alpha}}{C-Nc}$
$CNK_ikc^{1-\alpha}$
rescaled objective
c
10
9.5
9
8.5
8
7.5
7
6.5
6
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8



(a)



(b)

Figure 2: Plots of the rescaled objective functions $CK_ik\frac{c^{-\alpha}}{\lfloor\frac{C}{Nc}\rfloor-1}$, $CNK_ikc^{1-\alpha}$, and $CNK_ik\frac{c^{1-\alpha}}{C-Nc}$, for $\alpha=0.5$ (a), $\alpha=1.5$ (b), and $\alpha=1$ (c). The FEGLS case.

Numerical results

In this Section, the theoretical results obtained in the chapter are tested through simulations. For each c, the following empirical approximation of the summation of

the generalization error over all the units, conditioned on the training input data, is adopted. It is based on $\mathcal{N}^{tr}$ training sets and N_i^{test} test examples for each unit i $(i = 1, \ldots, N)$, hence on a total number $N^{test} = \sum_{i=1}^{N} N_i^{test}$ of test examples:

$$\sum_{i=1}^{N} \mathbb{E}\left\{ \left(\hat{\eta}_{i,FEGLS} + \underline{\hat{\beta}}'_{FEGLS} \underline{x}_i^{test} - \eta_i - \underline{\beta}' \underline{x}_i^{test} \right)^2 \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T} \right\}$$

$$\simeq \quad \frac{1}{N^{test}} \sum_{i=1}^{N} \sum_{h=1}^{N_i^{test}} \frac{1}{\mathcal{N}^{tr}} \sum_{j=1}^{\mathcal{N}^{tr}} \left(\hat{\eta}_{i,FEGLS}^{j} + \left(\underline{\hat{\beta}}_{FEGLS}^{j} \right)' \underline{x}_{i,h}^{test} - \eta_i - \underline{\beta}' \underline{x}_{i,h}^{test} \right)^2 . \quad (2.30)$$

In Eq. (2.30), $(\underline{x}_{i,h}^{test}, y_{i,h}^{test})$ is the h-th generated test example for the unit i, and $\underline{\hat{\beta}}_{FEGLS}^{j}$ is the estimate of the vector $\underline{\beta}$ obtained using the j-th generated training set. Similarly, $\hat{\eta}_{i,FEGLS}^{j}$ is the estimate of the individual constant η_i obtained using the j-th generated training set. For each choice of c, all the $\mathcal{N}^{tr}$ generated training sets share the same training input data matrices X_n, but differ in the random choice of the measurement noise. The number of rows in each matrix X_n is increased when c is reduced from $c_{\max}$ to $c_{\min}$, by increasing the number of observations T. For a fair comparison, when doing this, the rows already present in each matrix X_n are kept fixed. Finally, the same test examples (generated independently from the training sets) are used to assess the performance of the fixed effects generalized least squares estimates for different costs per example c.

For the simulations, we choose $N = 20$ for the number of units, $p = 5$ for the number of features, $c_{\min} = 2$, $c_{\max} = 4$, $\mathcal{N}^{tr} = 100$ for the number of training sets, $N_i^{test} = 50$ for the number of test examples per unit (hence the total number of test examples is $N^{test} = 1000$). The number of training examples per unit is $T = 50$ for $c = c_{\min}$, and $T = 25$ for $c = c_{\max}$. In this way, the (upper bound on the) total supervision cost is $C = 2000$ for both cases. Without loss of generality, the constant k in the model (2.27) of the variance of the supervision cost is assumed to be equal to 1. The components of the parameter vector $\underline{\beta}$ are generated randomly and independently according to a uniform distribution on $[-1, 1]$, obtaining

$$\underline{\beta} = [-0.8562, 0.6837, 0.2640, -0.0038, -0.0598]' . \quad (2.31)$$

Similarly, the fixed effects η_n (for $n = 1, \ldots, N$) are generated randomly and independently according to a uniform distribution on $[-1, 1]$, obtaining the vector

$$\underline{\eta} \quad = \quad \Big[-0.2330, -0.2779, -0.0434, -0.9707, 0.6848, 0.0720, -0.2033, -0.6877, 0.5967, -0.7895,$$

$$0.6500, 0.9717, 0.9673, -0.1443, -0.4211, 0.3109, 0.5189, 0.4709, 0.4414, -0.8382\Big]' \in \mathbb{R}^N \quad (2.32)$$

For both training and test sets, the input data associated with each unit are generated as realizations of a multivariate Gaussian distribution with mean $\underline{0}$ and covariance matrix

$$\mathrm{Var}\left(\underline{x}_{n,t}\right) = \mathrm{Var}\left(\underline{x}_i^{test}\right) = \begin{bmatrix} 1.4016 & 0.8086 & 1.2594 & 0.9866 & 0.6206 \\ 0.8086 & 0.9988 & 0.9518 & 1.2044 & 0.5003 \\ 1.2594 & 0.9518 & 1.9087 & 1.5945 & 0.7120 \\ 0.9866 & 1.2044 & 1.5945 & 1.9089 & 0.8294 \\ 0.6206 & 0.5003 & 0.7120 & 0.8294 & 0.4776 \end{bmatrix}, \quad (2.33)$$

which is symmetric and positive definite. This covariance matrix has been generated by setting $\mathrm{Var}\left(\underline{x}_{n,t}\right) = \mathrm{Var}\left(\underline{x}_i^{test}\right) = A_x A_x'$, where the elements of $A_x \in \mathbb{R}^{p \times p}$ have been randomly and independently generated according to a uniform probability density on the interval [0,1]. The parameter ρ in the covariance matrix (2.3) of the zero-mean vector of measurement noises (modeled in the simulations by a multivariate Gaussian distribution) is chosen to be equal to 0.3. As a robustness check, the whole procedure is repeated 100 times.

The results of the analysis are displayed in Tables 1 (for $\alpha = 0.5$), 2 (for $\alpha = 1.5$), and 3 (for $\alpha = 1$). Each table reports the results obtained in each repetition for $c = c_{\min}$ and $c = c_{\max}$. The total simulation time (for a MATLAB 9.4 implementation of the procedure) is of about 501 sec on a notebook with a 2.30 GHz Intel(R) Core(TM) i5-4200U CPU and 6 GB of RAM. A statistical analysis of the elements of the tables leads to the following results:

i) for $\alpha = 0.5$ (Table 1), the application of a one-sided Wilcoxon matched-pairs signed-rank test (Barlow, 1993, Section 9.2.3) rejects the null hypothesis that the difference between the approximated performance index from Eq. (2.30) for $c = c_{\max}$ and the one for $c = c_{\min}$ has a symmetric distribution around its median and that median is smaller than or equal to 0 (p-value=$1.9780 \cdot 10^{-18}$, significance level set to 0.05);

ii) for $\alpha = 1.5$ (Table 2), the application of a one-sided Wilcoxon matched-pairs signed-rank test rejects the null hypothesis that the same difference as above has

a symmetric distribution around its median and that median is larger than or equal to 0 (p-value=$1.9780 \cdot 10^{-18}$, significance level set to 0.05);

iii) for $\alpha = 1$ (Table 3), the application of a two-sided Wilcoxon matched-pairs signed-rank test fails to reject the null hypothesis that the same difference as above has a symmetric distribution around its median and that median is equal to 0 (p-value=0.4453, significance level set to 0.05).

Concluding, the following tables show that the simulation results are in perfect agreement with the theoretical ones, leading to the same conclusions. Interestingly, this holds even though relatively small values for T have been chosen for the simulations.

	(Repetition number) Approximated performance index from Eq. (2.30)				
$c = c_{\min}$	(1) $5.461 \cdot 10^{-4}$	(2) $6.127 \cdot 10^{-4}$	(3) $5.710 \cdot 10^{-4}$	(4) $5.638 \cdot 10^{-4}$	(5) $5.792 \cdot 10^{-4}$
	(6) $5.535 \cdot 10^{-4}$	(7) $6.084 \cdot 10^{-4}$	(8) $5.756 \cdot 10^{-4}$	(9) $5.976 \cdot 10^{-4}$	(10) $5.669 \cdot 10^{-4}$
	(11) $5.496 \cdot 10^{-4}$	(12) $5.562 \cdot 10^{-4}$	(13) $6.043 \cdot 10^{-4}$	(14) $5.762 \cdot 10^{-4}$	(15) $6.391 \cdot 10^{-4}$
	(16) $6.071 \cdot 10^{-4}$	(17) $6.007 \cdot 10^{-4}$	(18) $5.925 \cdot 10^{-4}$	(19) $5.600 \cdot 10^{-4}$	(20) $6.244 \cdot 10^{-4}$
	(21) $5.791 \cdot 10^{-4}$	(22) $5.590 \cdot 10^{-4}$	(23) $5.230 \cdot 10^{-4}$	(24) $5.339 \cdot 10^{-4}$	(25) $5.342 \cdot 10^{-4}$
	(26) $5.713 \cdot 10^{-4}$	(27) $5.824 \cdot 10^{-4}$	(28) $5.941 \cdot 10^{-4}$	(29) $5.727 \cdot 10^{-4}$	(30) $5.798 \cdot 10^{-4}$
	(31) $5.954 \cdot 10^{-4}$	(32) $5.819 \cdot 10^{-4}$	(33) $5.640 \cdot 10^{-4}$	(34) $5.779 \cdot 10^{-4}$	(35) $5.824 \cdot 10^{-4}$
	(36) $5.608 \cdot 10^{-4}$	(37) $5.565 \cdot 10^{-4}$	(38) $5.527 \cdot 10^{-4}$	(39) $5.981 \cdot 10^{-4}$	(40) $5.395 \cdot 10^{-4}$
	(41) $5.944 \cdot 10^{-4}$	(42) $6.110 \cdot 10^{-4}$	(43) $5.540 \cdot 10^{-4}$	(44) $5.490 \cdot 10^{-4}$	(45) $5.771 \cdot 10^{-4}$
	(46) $6.150 \cdot 10^{-4}$	(47) $5.492 \cdot 10^{-4}$	(48) $5.921 \cdot 10^{-4}$	(49) $5.552 \cdot 10^{-4}$	(50) $5.810 \cdot 10^{-4}$
	(51) $5.731 \cdot 10^{-4}$	(52) $6.018 \cdot 10^{-4}$	(53) $6.140 \cdot 10^{-4}$	(54) $5.836 \cdot 10^{-4}$	(55) $5.530 \cdot 10^{-4}$
	(56) $5.866 \cdot 10^{-4}$	(57) $5.661 \cdot 10^{-4}$	(58) $5.938 \cdot 10^{-4}$	(59) $5.795 \cdot 10^{-4}$	(60) $5.979 \cdot 10^{-4}$
	(61) $5.966 \cdot 10^{-4}$	(62) $5.882 \cdot 10^{-4}$	(63) $5.687 \cdot 10^{-4}$	(64) $5.718 \cdot 10^{-4}$	(65) $6.014 \cdot 10^{-4}$
	(66) $5.774 \cdot 10^{-4}$	(67) $5.872 \cdot 10^{-4}$	(68) $5.566 \cdot 10^{-4}$	(69) $5.678 \cdot 10^{-4}$	(70) $5.845 \cdot 10^{-4}$
	(71) $5.531 \cdot 10^{-4}$	(72) $5.446 \cdot 10^{-4}$	(73) $5.700 \cdot 10^{-4}$	(74) $6.055 \cdot 10^{-4}$	(75) $5.727 \cdot 10^{-4}$
	(76) $6.240 \cdot 10^{-4}$	(77) $5.616 \cdot 10^{-4}$	(78) $5.876 \cdot 10^{-4}$	(79) $6.031 \cdot 10^{-4}$	(80) $5.869 \cdot 10^{-4}$
	(81) $6.142 \cdot 10^{-4}$	(82) $5.764 \cdot 10^{-4}$	(83) $5.530 \cdot 10^{-4}$	(84) $5.901 \cdot 10^{-4}$	(85) $5.795 \cdot 10^{-4}$
	(86) $5.794 \cdot 10^{-4}$	(87) $5.818 \cdot 10^{-4}$	(88) $5.674 \cdot 10^{-4}$	(89) $5.512 \cdot 10^{-4}$	(90) $5.887 \cdot 10^{-4}$
	(91) $5.716 \cdot 10^{-4}$	(92) $6.050 \cdot 10^{-4}$	(93) $5.423 \cdot 10^{-4}$	(94) $5.883 \cdot 10^{-4}$	(95) $5.705 \cdot 10^{-4}$
	(96) $5.665 \cdot 10^{-4}$	(97) $5.732 \cdot 10^{-4}$	(98) $5.462 \cdot 10^{-4}$	(99) $5.896 \cdot 10^{-4}$	(100) $5.875 \cdot 10^{-4}$
$c = c_{\max}$	(1) $8.347 \cdot 10^{-4}$	(2) $8.193 \cdot 10^{-4}$	(3) $7.994 \cdot 10^{-4}$	(4) $8.132 \cdot 10^{-4}$	(5) $8.198 \cdot 10^{-4}$
	(6) $8.281 \cdot 10^{-4}$	(7) $7.627 \cdot 10^{-4}$	(8) $7.891 \cdot 10^{-4}$	(9) $8.281 \cdot 10^{-4}$	(10) $8.268 \cdot 10^{-4}$
	(11) $8.277 \cdot 10^{-4}$	(12) $8.097 \cdot 10^{-4}$	(13) $8.396 \cdot 10^{-4}$	(14) $8.187 \cdot 10^{-4}$	(15) $8.614 \cdot 10^{-4}$
	(16) $8.461 \cdot 10^{-4}$	(17) $8.299 \cdot 10^{-4}$	(18) $8.477 \cdot 10^{-4}$	(19) $8.171 \cdot 10^{-4}$	(20) $8.422 \cdot 10^{-4}$
	(21) $7.651 \cdot 10^{-4}$	(22) $8.068 \cdot 10^{-4}$	(23) $7.859 \cdot 10^{-4}$	(24) $8.034 \cdot 10^{-4}$	(25) $8.479 \cdot 10^{-4}$
	(26) $7.741 \cdot 10^{-4}$	(27) $7.839 \cdot 10^{-4}$	(28) $8.243 \cdot 10^{-4}$	(29) $7.620 \cdot 10^{-4}$	(30) $7.543 \cdot 10^{-4}$
	(31) $8.296 \cdot 10^{-4}$	(32) $8.280 \cdot 10^{-4}$	(33) $8.299 \cdot 10^{-4}$	(34) $8.115 \cdot 10^{-4}$	(35) $8.372 \cdot 10^{-4}$
	(36) $8.085 \cdot 10^{-4}$	(37) $8.362 \cdot 10^{-4}$	(38) $8.357 \cdot 10^{-4}$	(39) $8.585 \cdot 10^{-4}$	(40) $7.864 \cdot 10^{-4}$
	(41) $8.572 \cdot 10^{-4}$	(42) $8.098 \cdot 10^{-4}$	(43) $7.839 \cdot 10^{-4}$	(44) $7.941 \cdot 10^{-4}$	(45) $7.923 \cdot 10^{-4}$
	(46) $8.157 \cdot 10^{-4}$	(47) $8.743 \cdot 10^{-4}$	(48) $8.239 \cdot 10^{-4}$	(49) $8.181 \cdot 10^{-4}$	(50) $8.134 \cdot 10^{-4}$
	(51) $8.727 \cdot 10^{-4}$	(52) $8.600 \cdot 10^{-4}$	(53) $7.804 \cdot 10^{-4}$	(54) $8.078 \cdot 10^{-4}$	(55) $7.901 \cdot 10^{-4}$
	(56) $7.954 \cdot 10^{-4}$	(57) $7.811 \cdot 10^{-4}$	(58) $8.182 \cdot 10^{-4}$	(59) $8.339 \cdot 10^{-4}$	(60) $8.384 \cdot 10^{-4}$
	(61) $8.143 \cdot 10^{-4}$	(62) $8.129 \cdot 10^{-4}$	(63) $8.210 \cdot 10^{-4}$	(64) $8.319 \cdot 10^{-4}$	(65) $8.468 \cdot 10^{-4}$
	(66) $7.811 \cdot 10^{-4}$	(67) $8.211 \cdot 10^{-4}$	(68) $7.470 \cdot 10^{-4}$	(69) $8.128 \cdot 10^{-4}$	(70) $8.399 \cdot 10^{-4}$
	(71) $8.600 \cdot 10^{-4}$	(72) $8.537 \cdot 10^{-4}$	(73) $8.524 \cdot 10^{-4}$	(74) $8.117 \cdot 10^{-4}$	(75) $8.372 \cdot 10^{-4}$
	(76) $7.895 \cdot 10^{-4}$	(77) $8.114 \cdot 10^{-4}$	(78) $8.161 \cdot 10^{-4}$	(79) $8.537 \cdot 10^{-4}$	(80) $8.159 \cdot 10^{-4}$
	(81) $7.802 \cdot 10^{-4}$	(82) $8.178 \cdot 10^{-4}$	(83) $7.546 \cdot 10^{-4}$	(84) $7.922 \cdot 10^{-4}$	(85) $8.380 \cdot 10^{-4}$
	(86) $8.011 \cdot 10^{-4}$	(87) $8.541 \cdot 10^{-4}$	(88) $7.823 \cdot 10^{-4}$	(89) $8.026 \cdot 10^{-4}$	(90) $7.652 \cdot 10^{-4}$
	(91) $7.600 \cdot 10^{-4}$	(92) $7.859 \cdot 10^{-4}$	(93) $8.102 \cdot 10^{-4}$	(94) $8.599 \cdot 10^{-4}$	(95) $8.773 \cdot 10^{-4}$
	(96) $8.397 \cdot 10^{-4}$	(97) $8.105 \cdot 10^{-4}$	(98) $7.885 \cdot 10^{-4}$	(99) $8.061 \cdot 10^{-4}$	(100) $8.208 \cdot 10^{-4}$

Table 4: For $\alpha = 0.5$: values of the approximated performance index from Eq. (2.30) for the 100 repetitions of the simulation procedure.

(Repetition number) Approximated performance index from Eq. (2.30)					
$c = c_{\min}$	(1) $2.966 \cdot 10^{-4}$	(2) $3.008 \cdot 10^{-4}$	(3) $2.901 \cdot 10^{-4}$	(4) $2.950 \cdot 10^{-4}$	(5) $3.152 \cdot 10^{-4}$
	(6) $3.021 \cdot 10^{-4}$	(7) $2.611 \cdot 10^{-4}$	(8) $2.922 \cdot 10^{-4}$	(9) $2.814 \cdot 10^{-4}$	(10) $3.028 \cdot 10^{-4}$
	(11) $2.881 \cdot 10^{-4}$	(12) $2.937 \cdot 10^{-4}$	(13) $3.060 \cdot 10^{-4}$	(14) $3.077 \cdot 10^{-4}$	(15) $2.853 \cdot 10^{-4}$
	(16) $3.060 \cdot 10^{-4}$	(17) $2.952 \cdot 10^{-4}$	(18) $3.100 \cdot 10^{-4}$	(19) $2.876 \cdot 10^{-4}$	(20) $2.881 \cdot 10^{-4}$
	(21) $2.888 \cdot 10^{-4}$	(22) $3.084 \cdot 10^{-4}$	(23) $2.902 \cdot 10^{-4}$	(24) $2.902 \cdot 10^{-4}$	(25) $2.866 \cdot 10^{-4}$
	(26) $3.024 \cdot 10^{-4}$	(27) $2.866 \cdot 10^{-4}$	(28) $3.019 \cdot 10^{-4}$	(29) $2.921 \cdot 10^{-4}$	(30) $2.817 \cdot 10^{-4}$
	(31) $2.862 \cdot 10^{-4}$	(32) $2.828 \cdot 10^{-4}$	(33) $2.891 \cdot 10^{-4}$	(34) $2.842 \cdot 10^{-4}$	(35) $3.034 \cdot 10^{-4}$
	(36) $2.991 \cdot 10^{-4}$	(37) $2.870 \cdot 10^{-4}$	(38) $2.848 \cdot 10^{-4}$	(39) $2.837 \cdot 10^{-4}$	(40) $2.974 \cdot 10^{-4}$
	(41) $2.864 \cdot 10^{-4}$	(42) $2.724 \cdot 10^{-4}$	(43) $2.921 \cdot 10^{-4}$	(44) $2.991 \cdot 10^{-4}$	(45) $2.861 \cdot 10^{-4}$
	(46) $2.857 \cdot 10^{-4}$	(47) $2.887 \cdot 10^{-4}$	(48) $2.958 \cdot 10^{-4}$	(49) $2.985 \cdot 10^{-4}$	(50) $2.858 \cdot 10^{-4}$
	(51) $2.923 \cdot 10^{-4}$	(52) $2.698 \cdot 10^{-4}$	(53) $2.881 \cdot 10^{-4}$	(54) $3.008 \cdot 10^{-4}$	(55) $3.043 \cdot 10^{-4}$
	(56) $2.842 \cdot 10^{-4}$	(57) $2.781 \cdot 10^{-4}$	(58) $2.746 \cdot 10^{-4}$	(59) $2.819 \cdot 10^{-4}$	(60) $2.848 \cdot 10^{-4}$
	(61) $2.753 \cdot 10^{-4}$	(62) $3.010 \cdot 10^{-4}$	(63) $3.004 \cdot 10^{-4}$	(64) $2.805 \cdot 10^{-4}$	(65) $2.921 \cdot 10^{-4}$
	(66) $2.919 \cdot 10^{-4}$	(67) $2.947 \cdot 10^{-4}$	(68) $2.944 \cdot 10^{-4}$	(69) $2.960 \cdot 10^{-4}$	(70) $2.964 \cdot 10^{-4}$
	(71) $2.808 \cdot 10^{-4}$	(72) $2.940 \cdot 10^{-4}$	(73) $2.874 \cdot 10^{-4}$	(74) $2.851 \cdot 10^{-4}$	(75) $2.796 \cdot 10^{-4}$
	(76) $3.049 \cdot 10^{-4}$	(77) $2.885 \cdot 10^{-4}$	(78) $2.849 \cdot 10^{-4}$	(79) $2.711 \cdot 10^{-4}$	(80) $3.004 \cdot 10^{-4}$
	(81) $2.872 \cdot 10^{-4}$	(82) $2.908 \cdot 10^{-4}$	(83) $2.835 \cdot 10^{-4}$	(84) $2.779 \cdot 10^{-4}$	(85) $2.812 \cdot 10^{-4}$
	(86) $3.044 \cdot 10^{-4}$	(87) $2.736 \cdot 10^{-4}$	(88) $2.848 \cdot 10^{-4}$	(89) $2.815 \cdot 10^{-4}$	(90) $2.931 \cdot 10^{-4}$
	(91) $2.824 \cdot 10^{-4}$	(92) $2.923 \cdot 10^{-4}$	(93) $2.897 \cdot 10^{-4}$	(94) $2.872 \cdot 10^{-4}$	(95) $3.016 \cdot 10^{-4}$
	(96) $2.714 \cdot 10^{-4}$	(97) $2.807 \cdot 10^{-4}$	(98) $2.887 \cdot 10^{-4}$	(99) $2.838 \cdot 10^{-4}$	(100) $2.903 \cdot 10^{-4}$
$c = c_{\max}$	(1) $2.040 \cdot 10^{-4}$	(2) $2.029 \cdot 10^{-4}$	(3) $2.038 \cdot 10^{-4}$	(4) $2.021 \cdot 10^{-4}$	(5) $2.012 \cdot 10^{-4}$
	(6) $2.110 \cdot 10^{-4}$	(7) $2.030 \cdot 10^{-4}$	(8) $2.063 \cdot 10^{-4}$	(9) $2.064 \cdot 10^{-4}$	(10) $1.967 \cdot 10^{-4}$
	(11) $2.159 \cdot 10^{-4}$	(12) $2.019 \cdot 10^{-4}$	(13) $2.146 \cdot 10^{-4}$	(14) $2.027 \cdot 10^{-4}$	(15) $2.007 \cdot 10^{-4}$
	(16) $2.088 \cdot 10^{-4}$	(17) $1.979 \cdot 10^{-4}$	(18) $1.950 \cdot 10^{-4}$	(19) $2.023 \cdot 10^{-4}$	(20) $2.055 \cdot 10^{-4}$
	(21) $1.983 \cdot 10^{-4}$	(22) $2.081 \cdot 10^{-4}$	(23) $1.954 \cdot 10^{-4}$	(24) $2.213 \cdot 10^{-4}$	(25) $2.053 \cdot 10^{-4}$
	(26) $1.971 \cdot 10^{-4}$	(27) $2.031 \cdot 10^{-4}$	(28) $2.037 \cdot 10^{-4}$	(29) $1.976 \cdot 10^{-4}$	(30) $2.057 \cdot 10^{-4}$
	(31) $2.140 \cdot 10^{-4}$	(32) $2.043 \cdot 10^{-4}$	(33) $2.086 \cdot 10^{-4}$	(34) $2.087 \cdot 10^{-4}$	(35) $2.006 \cdot 10^{-4}$
	(36) $2.044 \cdot 10^{-4}$	(37) $1.967 \cdot 10^{-4}$	(38) $2.063 \cdot 10^{-4}$	(39) $1.953 \cdot 10^{-4}$	(40) $2.143 \cdot 10^{-4}$
	(41) $2.108 \cdot 10^{-4}$	(42) $2.105 \cdot 10^{-4}$	(43) $2.010 \cdot 10^{-4}$	(44) $1.970 \cdot 10^{-4}$	(45) $2.009 \cdot 10^{-4}$
	(46) $2.050 \cdot 10^{-4}$	(47) $1.948 \cdot 10^{-4}$	(48) $1.946 \cdot 10^{-4}$	(49) $2.093 \cdot 10^{-4}$	(50) $2.043 \cdot 10^{-4}$
	(51) $2.093 \cdot 10^{-4}$	(52) $2.036 \cdot 10^{-4}$	(53) $2.183 \cdot 10^{-4}$	(54) $2.022 \cdot 10^{-4}$	(55) $2.127 \cdot 10^{-4}$
	(56) $2.028 \cdot 10^{-4}$	(57) $2.020 \cdot 10^{-4}$	(58) $2.015 \cdot 10^{-4}$	(59) $2.028 \cdot 10^{-4}$	(60) $1.989 \cdot 10^{-4}$
	(61) $2.079 \cdot 10^{-4}$	(62) $2.199 \cdot 10^{-4}$	(63) $2.053 \cdot 10^{-4}$	(64) $2.127 \cdot 10^{-4}$	(65) $1.990 \cdot 10^{-4}$
	(66) $2.061 \cdot 10^{-4}$	(67) $1.983 \cdot 10^{-4}$	(68) $2.156 \cdot 10^{-4}$	(69) $2.073 \cdot 10^{-4}$	(70) $2.074 \cdot 10^{-4}$
	(71) $2.100 \cdot 10^{-4}$	(72) $2.024 \cdot 10^{-4}$	(73) $2.021 \cdot 10^{-4}$	(74) $1.989 \cdot 10^{-4}$	(75) $1.912 \cdot 10^{-4}$
	(76) $2.109 \cdot 10^{-4}$	(77) $2.043 \cdot 10^{-4}$	(78) $2.112 \cdot 10^{-4}$	(79) $2.015 \cdot 10^{-4}$	(80) $2.096 \cdot 10^{-4}$
	(81) $1.924 \cdot 10^{-4}$	(82) $2.071 \cdot 10^{-4}$	(83) $2.197 \cdot 10^{-4}$	(84) $2.173 \cdot 10^{-4}$	(85) $1.996 \cdot 10^{-4}$
	(86) $2.125 \cdot 10^{-4}$	(87) $1.978 \cdot 10^{-4}$	(88) $2.088 \cdot 10^{-4}$	(89) $2.011 \cdot 10^{-4}$	(90) $1.946 \cdot 10^{-4}$
	(91) $2.006 \cdot 10^{-4}$	(92) $2.156 \cdot 10^{-4}$	(93) $2.069 \cdot 10^{-4}$	(94) $2.018 \cdot 10^{-4}$	(95) $2.015 \cdot 10^{-4}$
	(96) $1.904 \cdot 10^{-4}$	(97) $1.983 \cdot 10^{-4}$	(98) $2.132 \cdot 10^{-4}$	(99) $1.933 \cdot 10^{-4}$	(100) $2.050 \cdot 10^{-4}$

Table 5: For $\alpha = 1.5$: values of the approximated performance index from Eq. (2.30) for the 100 repetitions of the simulation procedure.

(Repetition number) Approximated performance index from Eq. (2.30)					
$c = c_{\min}$	(1) $4.251 \cdot 10^{-4}$	(2) $4.155 \cdot 10^{-4}$	(3) $4.141 \cdot 10^{-4}$	(4) $4.162 \cdot 10^{-4}$	(5) $4.243 \cdot 10^{-4}$
	(6) $4.104 \cdot 10^{-4}$	(7) $4.018 \cdot 10^{-4}$	(8) $4.273 \cdot 10^{-4}$	(9) $4.224 \cdot 10^{-4}$	(10) $3.956 \cdot 10^{-4}$
	(11) $3.973 \cdot 10^{-4}$	(12) $4.068 \cdot 10^{-4}$	(13) $4.238 \cdot 10^{-4}$	(14) $4.102 \cdot 10^{-4}$	(15) $4.283 \cdot 10^{-4}$
	(16) $4.567 \cdot 10^{-4}$	(17) $4.224 \cdot 10^{-4}$	(18) $4.123 \cdot 10^{-4}$	(19) $4.362 \cdot 10^{-4}$	(20) $3.970 \cdot 10^{-4}$
	(21) $4.310 \cdot 10^{-4}$	(22) $4.298 \cdot 10^{-4}$	(23) $4.240 \cdot 10^{-4}$	(24) $4.399 \cdot 10^{-4}$	(25) $3.957 \cdot 10^{-4}$
	(26) $4.226 \cdot 10^{-4}$	(27) $4.144 \cdot 10^{-4}$	(28) $4.060 \cdot 10^{-4}$	(29) $4.025 \cdot 10^{-4}$	(30) $4.106 \cdot 10^{-4}$
	(31) $4.057 \cdot 10^{-4}$	(32) $4.060 \cdot 10^{-4}$	(33) $4.056 \cdot 10^{-4}$	(34) $4.183 \cdot 10^{-4}$	(35) $4.200 \cdot 10^{-4}$
	(36) $4.170 \cdot 10^{-4}$	(37) $3.990 \cdot 10^{-4}$	(38) $3.959 \cdot 10^{-4}$	(39) $4.103 \cdot 10^{-4}$	(40) $3.995 \cdot 10^{-4}$
	(41) $3.829 \cdot 10^{-4}$	(42) $4.041 \cdot 10^{-4}$	(43) $4.009 \cdot 10^{-4}$	(44) $3.815 \cdot 10^{-4}$	(45) $4.128 \cdot 10^{-4}$
	(46) $3.976 \cdot 10^{-4}$	(47) $4.249 \cdot 10^{-4}$	(48) $4.076 \cdot 10^{-4}$	(49) $4.253 \cdot 10^{-4}$	(50) $4.222 \cdot 10^{-4}$
	(51) $4.130 \cdot 10^{-4}$	(52) $4.011 \cdot 10^{-4}$	(53) $3.998 \cdot 10^{-4}$	(54) $4.047 \cdot 10^{-4}$	(55) $3.960 \cdot 10^{-4}$
	(56) $4.235 \cdot 10^{-4}$	(57) $4.157 \cdot 10^{-4}$	(58) $3.909 \cdot 10^{-4}$	(59) $4.221 \cdot 10^{-4}$	(60) $4.455 \cdot 10^{-4}$
	(61) $4.051 \cdot 10^{-4}$	(62) $4.077 \cdot 10^{-4}$	(63) $4.405 \cdot 10^{-4}$	(64) $4.106 \cdot 10^{-4}$	(65) $4.192 \cdot 10^{-4}$
	(66) $4.111 \cdot 10^{-4}$	(67) $4.183 \cdot 10^{-4}$	(68) $4.279 \cdot 10^{-4}$	(69) $4.099 \cdot 10^{-4}$	(70) $4.367 \cdot 10^{-4}$
	(71) $4.060 \cdot 10^{-4}$	(72) $4.016 \cdot 10^{-4}$	(73) $4.279 \cdot 10^{-4}$	(74) $4.080 \cdot 10^{-4}$	(75) $4.153 \cdot 10^{-4}$
	(76) $4.172 \cdot 10^{-4}$	(77) $4.084 \cdot 10^{-4}$	(78) $4.060 \cdot 10^{-4}$	(79) $4.187 \cdot 10^{-4}$	(80) $3.963 \cdot 10^{-4}$
	(81) $4.148 \cdot 10^{-4}$	(82) $4.097 \cdot 10^{-4}$	(83) $4.233 \cdot 10^{-4}$	(84) $3.991 \cdot 10^{-4}$	(85) $4.167 \cdot 10^{-4}$
	(86) $4.090 \cdot 10^{-4}$	(87) $4.176 \cdot 10^{-4}$	(88) $3.991 \cdot 10^{-4}$	(89) $4.027 \cdot 10^{-4}$	(90) $3.870 \cdot 10^{-4}$
	(91) $4.060 \cdot 10^{-4}$	(92) $4.177 \cdot 10^{-4}$	(93) $4.061 \cdot 10^{-4}$	(94) $4.133 \cdot 10^{-4}$	(95) $4.022 \cdot 10^{-4}$
	(96) $4.105 \cdot 10^{-4}$	(97) $3.803 \cdot 10^{-4}$	(98) $4.141 \cdot 10^{-4}$	(99) $4.171 \cdot 10^{-4}$	(100) $4.176 \cdot 10^{-4}$
$c = c_{\max}$	(1) $3.911 \cdot 10^{-4}$	(2) $4.069 \cdot 10^{-4}$	(3) $4.036 \cdot 10^{-4}$	(4) $4.297 \cdot 10^{-4}$	(5) $4.113 \cdot 10^{-4}$
	(6) $4.192 \cdot 10^{-4}$	(7) $4.082 \cdot 10^{-4}$	(8) $3.914 \cdot 10^{-4}$	(9) $4.029 \cdot 10^{-4}$	(10) $4.308 \cdot 10^{-4}$
	(11) $3.915 \cdot 10^{-4}$	(12) $3.739 \cdot 10^{-4}$	(13) $4.075 \cdot 10^{-4}$	(14) $4.111 \cdot 10^{-4}$	(15) $4.265 \cdot 10^{-4}$
	(16) $4.352 \cdot 10^{-4}$	(17) $3.934 \cdot 10^{-4}$	(18) $4.044 \cdot 10^{-4}$	(19) $4.112 \cdot 10^{-4}$	(20) $4.258 \cdot 10^{-4}$
	(21) $4.306 \cdot 10^{-4}$	(22) $4.179 \cdot 10^{-4}$	(23) $4.095 \cdot 10^{-4}$	(24) $4.189 \cdot 10^{-4}$	(25) $4.228 \cdot 10^{-4}$
	(26) $4.413 \cdot 10^{-4}$	(27) $3.976 \cdot 10^{-4}$	(28) $4.134 \cdot 10^{-4}$	(29) $4.166 \cdot 10^{-4}$	(30) $4.121 \cdot 10^{-4}$
	(31) $3.866 \cdot 10^{-4}$	(32) $4.440 \cdot 10^{-4}$	(33) $4.050 \cdot 10^{-4}$	(34) $4.129 \cdot 10^{-4}$	(35) $3.934 \cdot 10^{-4}$
	(36) $3.944 \cdot 10^{-4}$	(37) $4.066 \cdot 10^{-4}$	(38) $4.045 \cdot 10^{-4}$	(39) $4.115 \cdot 10^{-4}$	(40) $3.973 \cdot 10^{-4}$
	(41) $4.002 \cdot 10^{-4}$	(42) $4.248 \cdot 10^{-4}$	(43) $4.134 \cdot 10^{-4}$	(44) $4.302 \cdot 10^{-4}$	(45) $4.222 \cdot 10^{-4}$
	(46) $4.121 \cdot 10^{-4}$	(47) $3.946 \cdot 10^{-4}$	(48) $4.139 \cdot 10^{-4}$	(49) $4.183 \cdot 10^{-4}$	(50) $4.245 \cdot 10^{-4}$
	(51) $3.962 \cdot 10^{-4}$	(52) $4.204 \cdot 10^{-4}$	(53) $4.183 \cdot 10^{-4}$	(54) $3.930 \cdot 10^{-4}$	(55) $4.206 \cdot 10^{-4}$
	(56) $4.044 \cdot 10^{-4}$	(57) $3.754 \cdot 10^{-4}$	(58) $4.247 \cdot 10^{-4}$	(59) $4.185 \cdot 10^{-4}$	(60) $4.007 \cdot 10^{-4}$
	(61) $4.564 \cdot 10^{-4}$	(62) $4.174 \cdot 10^{-4}$	(63) $4.094 \cdot 10^{-4}$	(64) $3.944 \cdot 10^{-4}$	(65) $4.266 \cdot 10^{-4}$
	(66) $4.352 \cdot 10^{-4}$	(67) $4.042 \cdot 10^{-4}$	(68) $4.281 \cdot 10^{-4}$	(69) $4.168 \cdot 10^{-4}$	(70) $3.093 \cdot 10^{-4}$
	(71) $4.074 \cdot 10^{-4}$	(72) $4.007 \cdot 10^{-4}$	(73) $4.096 \cdot 10^{-4}$	(74) $3.968 \cdot 10^{-4}$	(75) $3.932 \cdot 10^{-4}$
	(76) $4.066 \cdot 10^{-4}$	(77) $4.213 \cdot 10^{-4}$	(78) $4.040 \cdot 10^{-4}$	(79) $4.300 \cdot 10^{-4}$	(80) $4.091 \cdot 10^{-4}$
	(81) $3.901 \cdot 10^{-4}$	(82) $4.161 \cdot 10^{-4}$	(83) $4.812 \cdot 10^{-4}$	(84) $4.039 \cdot 10^{-4}$	(85) $3.857 \cdot 10^{-4}$
	(86) $4.078 \cdot 10^{-4}$	(87) $4.267 \cdot 10^{-4}$	(88) $4.233 \cdot 10^{-4}$	(89) $3.985 \cdot 10^{-4}$	(90) $3.902 \cdot 10^{-4}$
	(91) $4.110 \cdot 10^{-4}$	(92) $4.045 \cdot 10^{-4}$	(93) $3.997 \cdot 10^{-4}$	(94) $4.170 \cdot 10^{-4}$	(95) $4.249 \cdot 10^{-4}$
	(96) $4.005 \cdot 10^{-4}$	(97) $3.942 \cdot 10^{-4}$	(98) $4.254 \cdot 10^{-4}$	(99) $4.215 \cdot 10^{-4}$	(100) $4.193 \cdot 10^{-4}$

Table 6: For $\alpha = 1$: values of the approximated performance index from Eq. (2.30) for the 100 repetitions of the simulation procedure.

Discussion and possible extensions

Up to our knowledge, the analysis and the optimization, made in the present Chapter, of the conditional generalization error in regression as a trade-off between training set size and precision of supervision, has been carried out only rarely in the literature. The reader is referred to Chapter 1 for a review of the correlated literature.

For what concerns practical applications, the theoretical results obtained in the anal-

ysis made in this chapter could be applied to the acquisition design of fixed effects panel data in both microeconometrics and macroeconometrics (Greene, 2003, Chapter 13). A semi-artificial validation on existing datasets could also be performed by inserting artificial noise with variance expressed as in Eq. (2.27), possibly with the inclusion of an additional constant term in that variance, to model the case of the original dataset before the insertion of the artificial noise. As a possible extension, one could investigate and optimize the trade-off between training set size and precision of supervision for the unbalanced FEGLS case (in which different units are associated with possibly different numbers of observations)[6], for the situation in which some parameters of the noise model have to be estimated either from the data or from a subset of the data, and for the case of a non-zero correlation of measurement errors for the observations associated with different units. Such developments could open the way to the application of the proposed framework to real-world problems, e.g., in econometrics. Another possible extension concerns the replacement in the investigation of the fixed effects panel data model with the random effects one (Greene, 2003, Chapter 13), which is also commonly applied to deal with the analysis of economic data, and differs from the fixed effects panel data model in that its parameters are random variables[7]. In the present analysis, however, a possible advantage of the fixed effects panel data model is that it also makes it possible to get estimates of the individual constants η_n (see Eq. (2.12)), which appear in the expression (2.16) of the conditional generalization error. Finally, another possible extension involves the case of dynamic panel data models (Cameron and Trivedi, 2005, Chapter 21).

[6]The unbalanced case in which all the measurement errors are uncorrelated - therefore, the FEGLS model is replaced in the analysis by the simpler Fixed Effects (FE) model - is the subject of our recent work Gnecco, Nutarelli, and Selvi (2020).

[7]If the additional assumptions of the random effects model hold, then both the fixed and the random effects estimates are consistent, but the latter is more efficient than the former. However, if they do not hold, then the random effects model provides inconsistent estimates (Greene, 2003, Chapter 13).

Part 2: Machine Learning and Econometrics

Chapter 3

Innovation and market size: an econometric application[1]

As reported in the Introduction of the present Dissertation, there are problems for which econometrics tools are still more robust and powerful than machine learning ones.

The present Chapter introduces one of such instances and specifically the estimation of the effect of market size on innovation in the Pharmaceutical industry.

Exploring the actual relationship between market rewards and innovation has been widely investigated in innovation economics for a long time (Scherer, 1982, Schmookler, 2013, Klepper and Malerba, 2010). This opened to the possibility of public demand in stimulating innovation, such as in the case of orphan drugs. Schmookler's "demand-pull" hypothesis, implying that innovation is a function of market demand, has been challenged over the years. Already in the '90s Kleinknecht and Verspagen (1990) noticed that the direction of causality between market size and innovation appears to be far from obvious. In particular, the authors suggested the presence of a simultaneous relationship between demand and innovation but did not manage to control for it. More recently, Stoneman (2010) and G. P. Ball, Shah, and Wowak (2018) developed more rigorous ways to detect such type of endogeneity.

Acemoglu and Linn (2004) developed a strategy to overcome the endogeneity bias at

[1]This chapter is partially based on *"Product recalls, market size and innovation in the pharmaceutical industry"* with M. Riccaboni

the market level. Specifically, they exploited changes in the market size for different drug categories driven by U.S. demographic trends (Acemoglu and Linn, 2004). After the contribution of Acemoglu and Linn, 2004, the focus moved from ascertaining the presence of the reverse causality of market size and innovation to detecting the best instrument for market size. Indeed, the instrument adopted in Acemoglu and Linn, 2004 was later criticized by Cerda (2007) as being itself endogenous. As detailed in Cerda (2007), while pharmaceutical innovation increases the age of patients, the fact that the average age increases imply that more patients would need innovative products. This scenario presents, again, the problem of reverse causality. Indeed, while demographic trends affect market size which have an impact of innovation, the latter influences, in turn, demographic trends.

To the best of our knowledge, such a gap in the literature is still unfilled.

Besides, authors, pushed by the studies of Acemoglu and Linn (2004), mainly concentrate their efforts on the Pharmaceutical industry. The Pharmaceutical, indeed, constitutes an ideal case study: in such an industry, consumers' needs are diverse and almost constant over time, which allows to separate it into independent sub-markets based on such needs (Bertoni et al., 2010). Furthermore, investments in innovation are vital for the industry's existence. Innovation is also relatively more easily measurable, being one of the major outputs of the Pharmaceutical industry. In the Pharmaceutical industry, market size is defined based on the Anatomical Therapeutic Chemical (ATC) Classification System, i.e., a drug classification system classifying *"the active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties"* (Skrbo, A. et al. (2004)).

The present Chapter captures the relationships among market size and innovation at the ATC-3 level by instrumenting market size with recalls (see below) of drugs operated by the Food and Drug Administration (FDA). The Chapter contributes to the literature in three ways.

First, we adopt an innovative measure of innovation, i.e., the overall number of trials at the ATC-3 level instead of the cumulative R&D expenditures or New Molecular Entities (NME). This necessary modification overcomes the limitations of the other two most adopted measures. R&D expenditures are, indeed, linked both to firms' long-term profit decisions (Cohen, 2010) and, more critically, to their size. Stone-

man (2010), among others, suggested that smaller entrants might be more inclined to invest in R&D expenditures than their bigger veteran competitors. Hence, innovation as cumulative R&D expenditures of the firms composing the market might be related to market size as measured by the firms' cumulative sales composing the market. More delicate is the topic concerning NME, also adopted in Acemoglu and Linn (2004). NME are innovative products that contain active moieties, i.e., significant parts of the molecule. Such molecule parts were not approved by the FDA previously. They can be innovative products that have never been exploited in clinical practice. Alternatively, they can also be related to previously approved products. Though a complete definition, the one of NME does not fully capture, in our opinion, the will of innovation by firms inside the market. The reason hinges around the stage of drug approvals at which NME, with respect to the measure of innovation employed in the present work, are approved by FDA.

Pharmaceutical drug approval is a long process. Firms should first pass a pre-clinical phase, a stage of research that starts before that clinical trials (testing in humans) can begin. During the pre-clinical phase, public agencies collect critical data on the feasibility of the trial, iterative testing, and safety of the tested drug. The clinical drug development stage, then, consists of three phases. In Phase 1, the company conducts clinical trials on healthy individuals. This process helps determining the drug's basic properties and safety profile in humans. According to DiMasi, R. W. Hansen, and Grabowski, 2003, *"typically, the drug remains in this stage for one to two years"*. Once the drug is dispensed to volunteers of the targeted population, Phase 2 begins. The latter consists, basically, on performing the trial on a larger group of individuals with respect to the previous Phase. Finally, Phase 3 compares a new drug to a standard-of-care drug. NME are FDA-approved entities having overcome pre-clinical trials. The number of trials, adopted in the present Chapter, also considers the pre-clinical phase. In other words, the latter measure takes also into account potentially unsuccessful trials, i.e., trials not passed to the clinical phase, which equally characterize an innovative drive of the firm. A second contribution is given by the adoption of a more refined classification of the relevant market segments, i.e. ATC-3 classes. The available data on the ATC-3 level of classification well captures the structure of sub-markets, usually constructed artificially or disregarded by the literature. The lit-

erature, in fact, mostly uses ATC-2 and ATC-1 level. When ATC-3 level is employed is mainly for sake of comparison with other broader levels.
Moreover, the ATC-3 level is the level employed by antitrust agencies. The reader is referred to Section 2 for details.
A further improvement is methodological. The Chapter adopts an IV approach to deal with the endogeneity problem of market size. The enhancement compared to past research consists of the instrumentation of market size with recalls to overcome the endogeneity issue already detailed. The idea is to exploit sharp and unexpected recalls. The task of characterizing recalls as being sharp and unexpected requires a general definition of recalls. We believe that recalls are exogenous. The idea is that, at the market level, firms cannot anticipate a recall being issued to a competitor. In Section 3 we provide several arguments in favor of the sharpness of recalls not being necessarily issued in "more risky" ATC markets.
FDA refers to a recall as *"the most effective way to protect the public from a defective or potentially harmful product. A recall is a voluntary action taken by a company to remove a defective drug product from the market. Drug recalls are conducted either on a company's initiative or by FDA request"* (FDA U.S. Food Drug, 2019). In a recall, the FDA's role is to oversee a company's strategy, assess the recall's adequacy, and classify the recall. According to their severity, the FDA classifies the recalls in Class I (more severe), Class II, and Class III (least severe). Medicines may be recalled for several reasons ranging from health hazards to potential contamination, adverse reaction, mislabeling, and poor manufacturing. Recalls should not be confused with withdrawals. Unlike the FDA definition, literature often refers to withdrawals as post-marketing recalls imposed by the FDA on firms due to their high severity and risk to human health. Therefore, recalls can be expected and voluntarily made by firms if minor or sharp and unexpected if most severe and forced by the FDA. In other words, according to the definition often adopted in the literature, withdrawals are post-marketing recalls and are often operated on after a severe Class I recall. Following Onakpoya, I.J. et al. (2016), though there are discrepancies among counties, in the 72% recall procedure due to adverse effects ended up in a withdrawal [2]. To

[2]This estimation was manually computed following the list provided in Onakpoya, I.J. et al. (2016)

be consistent with our recall data, looking at major recalls, the majority of Class I recalls containing the word "death" within their root cause resulted in a withdrawal. By the very definition of drug recalls, we expect a drop in sales consequent to a drug recall in a market. To clarify the latter mechanism, one can refer to Merck's popular recall of VIOXX in 2004. VIOXX was withdrawn from the market due to an increased risk for serious cardiovascular events. The recall caught unprepared both the market and the firm. After the announcement of the recall of VIOXX in September 2004, shares of Merck and its sales dropped. This drop was publicized by mass media (Terence N., 2004, Bowe C., 2005 among others) and well recognized by academics (see, e.g. C. H. Tong, L.-I. Tong, and J. E. Tong, 2009 among others).

In the present Chapter we try to assess such sharp and unexpected recalls ("major recalls" from now on). The definition of major recalls we adopted throughout the Chapter has been recovered by filtering the causes of Class I recalls. We filtered recalls according to the relevance of the cause, its severity in terms of potential danger against human life, and the FDA's actions. Specifically, we comprised in the definition of major recalls, withdraws, Class I recalls containing critical keywords among their causes such that:"contamination," "death/s," "overdose," "symptoms," "particulate matters," and "adverse reaction".

We did not employ Class II recalls since, in our opinion, they constitute a weaker instrument than major recalls. Nonetheless, in the dedicated Appendix, we included the analysis using all types of recalls as a robustness check. The main results are confirmed.

1 Literature Review

The literature has acknowledged the importance of market size in explaining the rate of innovation for many years. Back in 1942, Schumpeter indicated that larger firms are more innovative than smaller ones. In the early '60s, the focus shifted more broadly on the possible effects of demand on market size (see, e.g., Scherer, 1982). It was not yet clear whether the reverse causality of demand and innovation played a relevant role. Scherer (1982), for instance, argued that causality ran primarily from

sales to innovation. The study of the author, however, has been criticized in several aspects. The definition of demand was indeed still too broad and was not conclusive about the unique sign of the relationship between demand and innovation, i.e. reverse causality (see, e.g., Mowery and Rosenberg, 1979). At the time, the research did not focus specifically on the pharmaceutical sector nor looked at the aggregate market level (see, e.g., Pakes and Schankerman, 1984).

Most recently, Kleinknecht and Verspagen (1990) denounced a clear reverse causality of demand and innovation, thus invalidating the prior studies. Geroski and Walters (1995) empirically verified such conclusions soon after, finding out how innovations increase demand by creating their demand.

Besides, it was clear that heterogeneous shifts of demand played a prominent role in determining technological development (see, e.g., Malerba, 2007). Between 1980 and 1990 and most recently in 2002, several studies showed, for instance, how innovation reacted elastically to energy prices.

Nowadays, a huge part of the research on the relationship between market size and innovation regards the pharmaceutical industry, where innovation represents a pushing power. Literature mainly takes into account two levels of aggregation: firm-level and market-level. Past research efforts have been devoted to identifying the impact of firm size on R&D investments and output. Nevertheless, this question is still an open debate (see Mellahi and Wilkinson, 2010, Kolluru and Mukhopadhaya, 2017 among others). Specifically, controversial results emerge due to the difficulty in fully excluding unobservable endogeneity sources varying with time. Such unobservables might derive from strategic decisions taken within the firms, which, in turn, might be related to their size. For example, small pharmaceutical firms are likely to take more risky decisions than big established ones (B. H. Hall and Rosenberg, 2010). Moving to market aggregation easily avoids the mentioned concerns. Unobservables related to market size can principally be considered as intrinsic characteristics of markets and, consequently, fixed in time. Thus, fixed effect techniques allow researchers to control for unobservable heterogeneity, purging the idiosyncratic endogeneity of market size. Therefore, the market seemed a more suitable level, and most authors shifted to the latter level of aggregation.

The literature on the pharmaceutical sector is vast. Part of its variability is due to

the measures of innovation adopted. Some authors adopted accounting data focusing on R&D. While R&D is robust under perfect capital markets, it becomes inconclusive with imperfect markets where current investment choices reflect the future ones. Within market imperfection, current revenues (market size) are a reasonable proxy for future market size. This, at first, might cause endogeneity problems due to the correlation of current revenues with unobservables (e.g. risk propensity of the firms' management). Moreover, since present R&D may be responding both to present and future sales opportunities, the coefficient of market size might incorporate two effects that are difficult to separate. Aware of such an issue, the authors included lagged proxies of the market size (see, e.g., Giaccotto, Santerre, and Vernon, 2005 who estimated that a 1% increase in price leads to a 0.58% increase in R&D spending). Other problems related to R&D measure are reported in B. H. Hall and Rosenberg, 2010. To give an idea, several authors (see e.g. Pammolli, Magazzini, and Riccaboni, 2011, Baumann and Kritikos, 2016 among others) showed how, though pharmaceutical firms made substantial investments in R&D, the latter did not produce innovation [3].

Other measures of innovation include clinical trials (see Kyle and McGahan, 2012 among others) and changes in Medicare part D. Medicare part D is an optional United States program to help beneficiaries of the national health insurance Medicare, to pay for self-administered prescription drugs. The usage of Medicare part-D as innovation measure might affect both present and future market size (Blume-Kohout and Sood, 2013, Dubois et al., 2015). As suggested in Blume-Kohout and Sood, 2013, *"Medicare Part D could have affected firms' R&D expenditures both to its expansion of expected future markets for products still in pipeline, and also via two supply side mechanisms."* (see Blume-Kohout and Sood, 2013 for further details). Thus, Medicare part-D might have stimulated current and, critically, future sales through R&D expenditures. Scholars found a positive response of innovation to shocks in market size. Again, the problem remained the possible co-occurrence in innovation's response to both current and expected cash flows generated by market size shocks. Within the estimated coefficient of market size was therefore impossible to distinguish between current and future effects. In the present chapter we mitigate such an issue by taking

[3]Thus undermining the validity of R&D investment as a good proxy for innovation.

all active projects having generated innovation and failed projects in all the clinical and pre-clinical Phases. Thus, our innovation measure affects only current market size. The future market size is, indeed, just partially influenced by currently active (and failed) projects. The latter hypothesis has been tested in the robustness checks (where the lag of innovation has been inserted as a covariate).

Besides, innovation has been quantified by the number of relevant medical journal articles (Lichtenberg, 2006). Further measurements comprise the number of new drugs launched, including generic drugs (Acemoglu and Linn, 2004, Dubois et al., 2015) in the form of New Molecular Entities (NME), New Chemical Entities (NCE), or approvals of new medicines by the FDA.

Similarly, many measures of market size have been embraced.

Acemoglu and Linn (2004) gave a first significant contribution on the relation between market size and innovation in the pharmaceutical industry. Their idea relies on adopting demographic shifts to instrument market size controlling for observables or unobservables arising from reverse causality. In particular, Acemoglu and Linn (2004) exploited variations in the expenditure share of different U.S. age cohorts for different therapeutic classes from 1970-2000. They discovered that a 1% increase in the shares of expenditure would lead to a rise of 4% in the number of new medicines, a far higher elasticity than the average elasticity found in the remaining literature (Dubois et al., 2015). Cerda (2007) provided further insights on the results found in Acemoglu and Linn (2004). Employing U.S. demographic data, Cerda (2007) showed that there are essential feedback effects not considered in Acemoglu and Linn (2004). New drugs might affect the market size through their impact on the mortality rate. Indeed, innovative medicines are likely to cure more diseases, raising the population's average age and, hence, the number of older people needing such cures. Demand shifts accordingly, bringing out again the issue of reverse causality.

Recent literature on the topic improves above all on the methodological part (see e.g. Lichtenberg, 2006, Civan and Maloney, 2009, Dubois et al., 2015, Rake, 2017 and others). Dubois et al., 2015's novelty pertains the usage of global pharmaceutical data at the ATC-1 and 2 levels. Dubois et al., 2015 still employs demographic shifts as instrument for market size. Civan and Maloney, 2009 found out that *"the higher the prices of existing drugs in a therapeutic category, the larger the number of drugs*

in the development pipeline in that therapeutic category". It is important to notice that Civan and Maloney, 2009's work is more focused in estimating the elasticity of drug development to the market price of drugs. Their results are conducted using R&D as a measure of innovation and suffer from the endogeneity of prices in the elasticity equations adopted to reach the results. Finally, Rake, 2017 adopts a unique database and a Poisson Quasi Maximum Likelihood approach to reach his results. His contribution is also related to the usage of New Molecular Entities (NME) and New Drug Approvals (NDA) as measures of innovation.
Authors found, on average, that a 1% increase in the market size measure increases innovation of 0.4% to 0.7%.
Past papers acted mainly at disease level or, at most, at ATC-1 or ATC-2 levels (see e.g.Dubois et al., 2015). To the best of our knowledge, no works are focusing on the more interesting ATC-3 level at which antitrust authorities work. According to us, there are several advantages of using drug classes rather than disease classes. Firstly, since firms make directly the request for undergoing a New Clinical Trials (NCTs), NMEs, or NDAs, devoting too much attention to the demand-side might neglect the supply-side dynamics, which induce firms to undergo an NDA, NCT or NME. In particular, aggregate sales of drug classes align to the supply-side dynamics, while sales based on disease classes (i.e., aggregated sales of products purchased by patients) are more related to the demand side.
In other words, while firms might follow demand-side stimuli to undergo an NDA (or an NCT), they, above all, look up at the competitors, i.e., products of other companies in the same ATC class. The latter applies to commercial trials when the sponsor is a pharmaceutical industry and not academy/research related. Thus, aggregating sales into disease classes, forces them to be clustered in a demand-driven group which is less informative on the supply-side dynamics (i.e. behavior of competitors) leading the firm to undergo an NME, NDA or NCT. When estimating the effect of market size grouped into disease-classes (demand-riven) and innovation (supply-driven), one might incur in an underestimation of the latter effect.
Furthermore, by taking disease classes, one includes in the definition of innovation different chemical and therapeutic typologies of drugs ranging from topical to systemic drugs, from vaccines to ointment. This lack of distinction might lead to endogeneity

through several channels, such as people's expectations. Patients might beware of some drugs, affecting the probability of having a larger market size for the product's typology under question. Other endogeneity sources regard the possibility of a correlation between regressors and the error term (which includes "drug-type"). For instance, regulations may be product type-specific (e.g., the regulations of the WHO vaccines do not apply to other drug types). Other possibly problematic controls are knowledge stocks, which could again depend on the product type. Moreover, knowledge stocks might increase by developing innovative medicines in classes where only a particular type of medicine has been developed until that moment. An example is provided in dermatology, where academics produce papers for adopting topical medicines for systemic usage due to some systemic medicines' undesired side effects.

Finally, the length of a clinical trial varies depending on the type of medicine under study, which may cause lagged effects of market size if disease class is employed.

To the best of our knowledge, among the several innovation measures, no work exploited INDs and early stages clinical trials (i.e., pre-clinical and Phase I) together with Phase II and Phase III trials.

The two more recent estimates of the relationship between market size and innovation have been provided in Rake (2017) and Dubois et al. (2015). The latter used NCE to measure innovation and defined market size as a measure of expected revenue. The dataset comprised information about sales for 14 different countries. Specifically, Dubois et al. (2015) measured market size as the total revenue over the entire life cycle of a branded drug. Dubois et al. (2015) performed a control function approach and recovered an estimate of the relation between market size and innovation for each therapeutic class at level 1. The average elasticity of innovation to the market size in Dubois et al. (2015) was about 23%, which is relatively low than the average estimates. A possible explanation can be found in Blume-Kohout and Sood (2013), which states that several of the countries chosen for the analysis regulate prescription drug prices, and regulations may change rapidly over time. Thus, given the lower expected profit per consumer and more significant uncertainty about future profits and prices, firms' R&D decisions are likely to be less responsive to a unit change in expected revenues for all these countries combined versus the exact unit change in the U.S. market.

Finally, Rake (2017) adopted several measures of innovation from NCE to clinical trials in Phase II and Phase III. Rake (2017) found no evidence of reverse causality when adopting NCE. One of his efforts was to account for changes in the industry's R&D process, from "random screening" to "guided drug development". (Rake, 2017) modeled technological opportunities and inserted them as regressors in the analysis, finding a positive relationship with Phases II and III trials. His results are in line with Cerda (2007) and Acemoglu and Linn (2004). Tab. 23 in the Appendix provides a schematic literature review on previous estimates of the relation between innovation and market size.

2 Data

The sales data employed come from Evaluate dataset. The controls have been extrapolated from Evaluate, from the PHarmaceutical Industry Database (PHID) and FDA. Specifically, some of the regressors derive from an elaboration of the variables present in the PHID database.
Sales data for the US pharmaceutical market range from 2004 to 2015. Sales data were initially available at the product and molecule level and have successively been aggregated at the ATC-3 level. In the ATC classification system, drugs are classified at five levels (ATC-1, ATC-2, ATC-3, ATC-4, ATC-5): the higher the level, the more detailed the classification. Acemoglu, Linn (2004) employed ATC-1 and ATC-2 categorizations to define market size. In particular, Acemoglu, Linn (2004) constructed market size as the *"the average expenditure share of drugs in (ATC-1 or ATC-2) category c in the total income of those in age group a"*.
The data at our disposal allow us to catch the diverse strata of products inside broader classes (ATC-1 and ATC-2) in terms of both demand and supply dynamics. Medicines classified inside an ATC-1 or an ATC-2 level can satisfy patients with completely diverse needs since they are designed to cure various diseases. Simultaneously, a firm investing in the same ATC-2 sector might invest in more ATC-3 sectors. In the case of ATC-1 or ATC-2 adoption, the missing information about the firms' investments in ATC-3 classes may lead to the construction of uninformative innovation and market size variables. Such variables might not consider the firms' specialization in a

sub-sector rather than in another one belonging to the same ATC-2 or ATC-1 class. In the Appendix, we have also evaluated other levels of analyses (firm, product, and ATC-firm aggregations) but opted for the ATC-3 level because of the importance of the ATC-3 level being employed by antitrust agencies. We avoid adopting the ATC-4 level since, at such a level of granularity, products belonging to a specific ATC-4 class might not differ substantially from others belonging to another ATC-4 class. This might lead to between-group dependencies (e.g., innovations in an ATC-4 may also affect a close ATC-4 class) which could cause inference to be invalid. Further, at the ATC-4 level, compensations may also intervene between groups, thus invalidating the strength of the instrumental variable recalls. Such compensations might regard, for instance the drop in sales due to a recall of a product belonging to an ATC-4 level with the sales another one belonging to another ATC-4 level. This is due to the high similarity of ATC-4 classes.

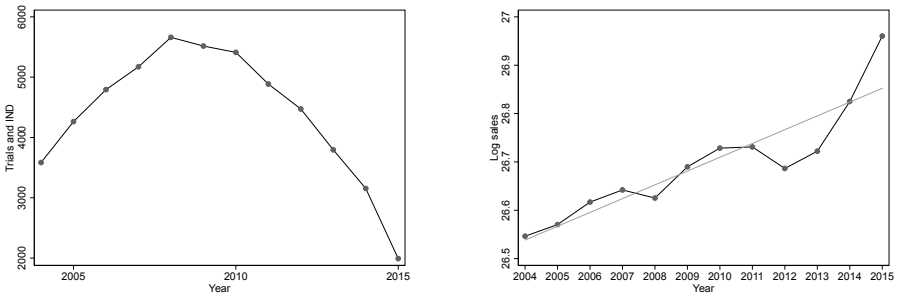
The available data also contain the launch date and ATC code of products. We focused on worldwide sales of US companies.

Data on NCT for 2004-2015 at product level come from the ClinicalTrials.gov website, while data on commercial Investigational New Drugs (IND) at product level derive from a Pharmaceutical Industry Database maintained at IMT Lucca.

Clinical trials are research studies performed on people who aim to evaluate a medical, surgical, or behavioral intervention (NIH, National Institute on Aging, 2020). An IND in clinical trials means a company working in the pharmaceutical industry is authorized to start human clinical trials. With an IND, the firm is also allowed to ship an experimental medicine across state lines before approving a marketing application.

Clinical trials comprise trials from Phase I to Phase IV. Fig.3 displays the yearly number of trials and commercial IND as obtained by the mentioned sources.

It also shows the expected positive trend of sales of the Pharmaceutical industry in time.



Number of trials and INDs per year Trend of sales by year

Figure 3: Overview of sales' and trials' trends.

A considerable drop in Trials and IND occurred after 2013, as it is evident from Fig.3 (a). The reason for such lack is that, in general, clinical trials innovate drugs, approaches, and interventions. However, approaches and interventions are excluded from the count of trials to focus strictly on innovation coming from industrial sources.

Recalls data have been manually collected from different sources, among which FDA website, openFDA, various articles, and web sources (e.g., Onakpoya, Heneghan, and Aronson, 2016; WHOCC website, PubMed, Siramshetty et al., 2016 and others).
The 7.19% of the firms' sample (i.e., 697 firms in total) have issued Class II recall. Among the firms that issued a recall, 51 firms underwent a recall of Class I, 27 of which issued a single recall of Class I, and just three firms issued more than 9 Class I recalls.
Besides, the recalls of pure compounders were only partially included [4] and, when included, were attributed to the unique manufacturer/distributor in the database. Finally, recalls coming from repackaging firms were not attributed uniquely to the repackager (e.g., Aidapak) but to the labeler specified in the NDC.
Due to the need of excluding recalls coming from repackagers and compounders, it is not easy to establish a unique and unambiguous pattern of recalls over the years. The situation is further complicated wherever different resources employ different methodologies to count the recalls. An example of the cited uncertainty in sources is found when comparing Unger Consulting (2017)[5] and Laguna Treatment Hospital (2019).

[4] Due to the unavailability of data. We verified that the representativeness of recalls is preserved F., 2021

[5] n.d.

Specifically, Unger Consulting (2017) asserts that the number of recalled products had remained reasonably constant except for 2010 and 2013, when the number went down by approximately 35%. The statement contrasts with what was reported in Laguna Treatment Hospital (2019). According to Laguna Treatment Hospital (2019) *"a spike in the number of drugs recalled occurred in 2013. There were nearly 60 recalls in that year alone. However, 2017, with 71 recalls, saw nearly the highest number of recalls since 2009. Only 2011 and 2009 surpassed it at 74, and 75 recalls, respectively".*
Tab. 7, provides a list of the primary sources and the average number of recalls across them. The following is an attempt to overcome the mentioned issues involving the dissimilarities of data origins.

Year	CNN	Regulatory Focus	K. Hall et al., 2016	FDA Enforcement Reports	Laguna Treat. Hospit.	AVERAGE
2004		68				68
2005		140				140
2006	384	109				243
2007	391	56				189
2008	426	128		176		244
2009	1742	85		1660		890
2010		135		389		262
2011		236		1279	75	530
2012		381	499	1518		799
2013		1031	1283	848	60	805
2014		640	1344	893		959
2015				1584		1584

Table 7: Sources with reported number of recalls

To overcome the dissimilarities of data origins, we chose the average as the benchmark to compare with the collected recalls. Fig.4 illustrates in more detail the comparison between the benchmark recalls represented by the average recalls among all sources and the collected recalls. The recalls of the repackager Aidapak of 2011 are considered as "outliers" and, for this reason, are not included among the collected recalls at this stage.

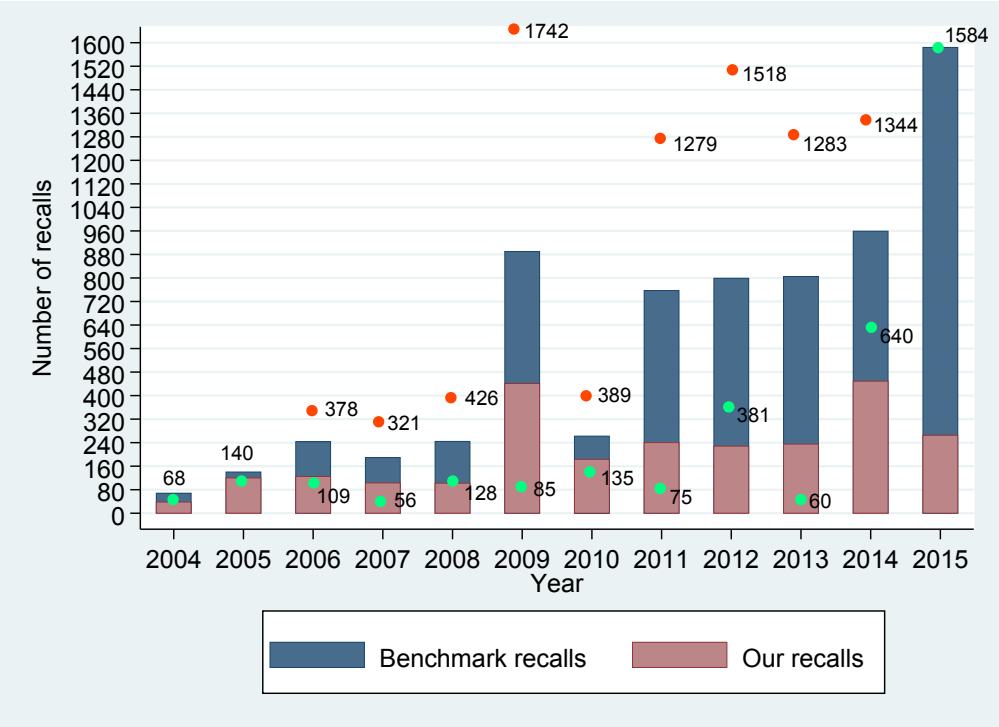


Figure 4: The number of our recalls against the number of the benchmark recalls (average of sources). We include the minimum and the maximum number of recalls retrieved by the different sources. Mint colored points represent the minimum amount of recalls retrieved among all the sources at our disposal. Red points represent the maximum number of recalls among all the sources. A single mint point has been put whenever a single source was present for a year (2004, 2005, 2015).

Fig. 4 underlines a disproportion in terms of the number of recalls starting from 2009, with respect to the benchmark. Such deficiency pertains to the counting methodology together with the structure of the database (see above).

Though the global trend is approximately reproduced, 2011, 2013, and 2015 represent problematic years due to the abundance of compounding an repackager recalls. The dissimilarity of 2011 concerning the benchmark can be easily explained. Indeed, with the exclusion of Aidapak's recalls from the count of the collected recalls, the latter dropped. Furthermore, 2013 and 2015 have far fewer recalls than expected because more than 60% of the recalls in 2013 and nearly 75% of the recalls in 2015 were represented by compounding firms.

The trend of recalls of the benchmark seems to be well reproduced, however, when the recalls of pure compounders are excluded from both the benchmark number and the sample of collected recalls. Fig. 5 shows, indeed, an accordance in trends. Aidapak's

recalls are here included. Indeed, Aidapak is a repackager and not a compounder. Besides, we want to show that 2011 does not constitute a problematic year once Aidapak's recalls are considered.

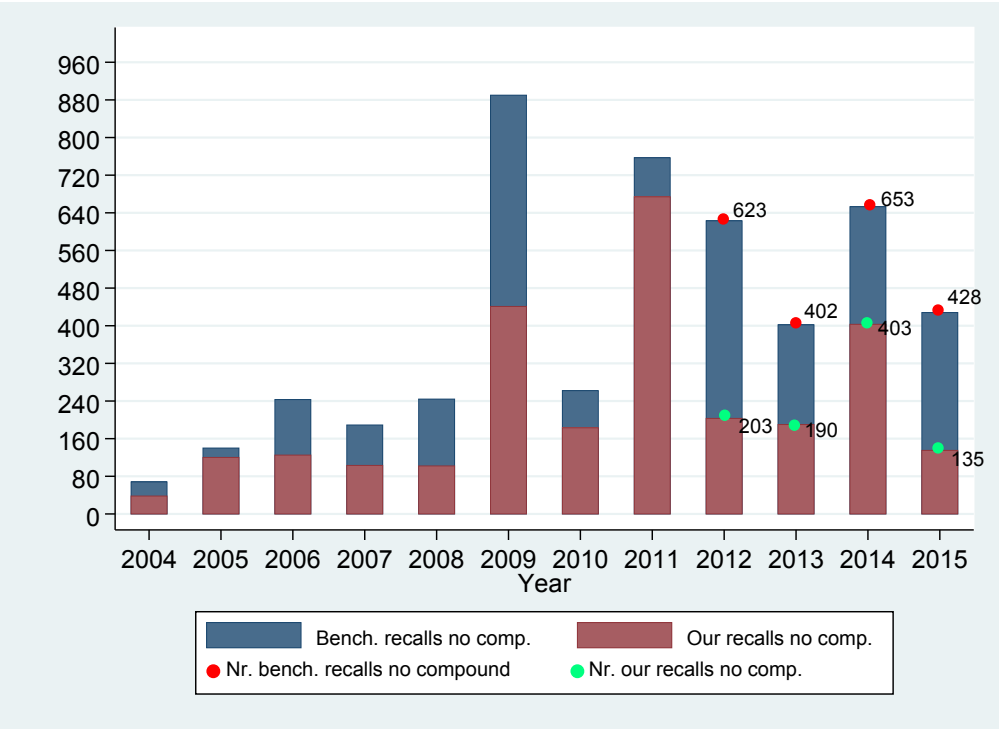


Figure 5: The number of our recalls against the number of the benchmark recalls without compounding recalls. We included the minimum and the maximum number of recalls retrieved by the different sources. Mint colored points represent the minimum amount of recalls retrieved among all the sources at our disposal. Red points represent the maximum number of recalls among all the sources.The situation i almost unchanged with respect to Fig. 4 until 2011. From 2011 on, the recalls collected in our dataset follow the benchmark if compounders' recalls are excluded more precisely.

To conclude, as a check of the exogeneity of recalls, we constructed a box plot displaying the average number of trials (and their dispersion) in both ATC markets having undergone a recall and not having undergone a recall by year. The latter exercise helps in understanding that major recalls do not necessarily intervene in more innovative markets. Indeed, Fig. 6 displays that the yearly number of trials of ATC markets undergoing major recalls almost coincides with the average number of trials in all other markets.

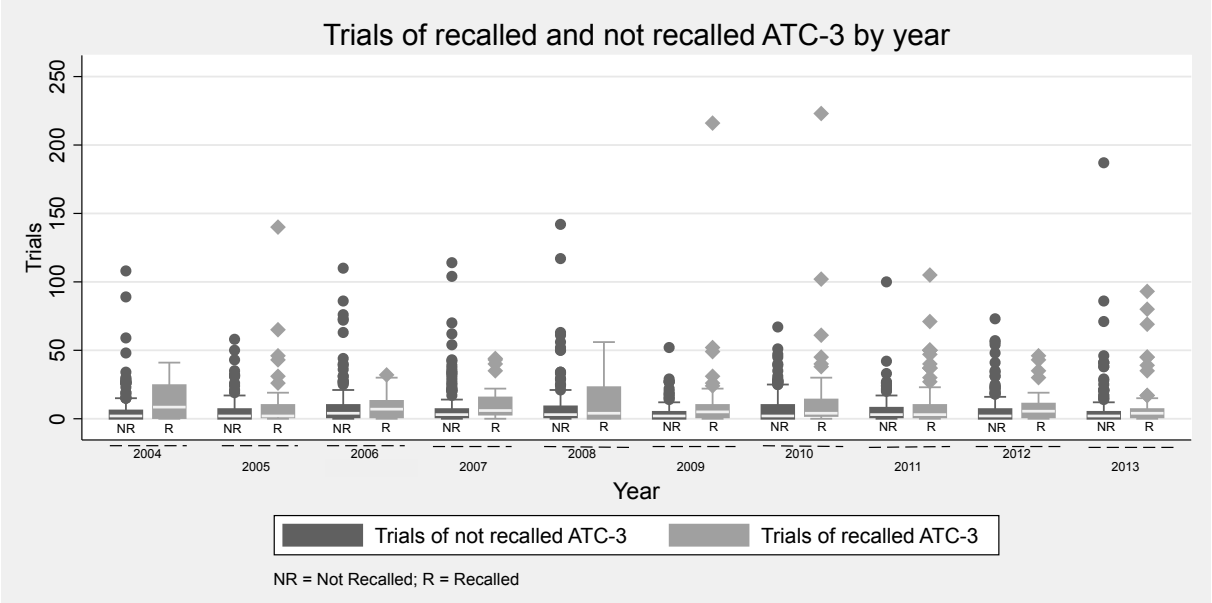


Figure 6: The box plots show how, on average, recalls do not necessarily happen in more innovative markets. The average number of trials amounts to 123 for ATC-3 markets having undergone a recall and 118 for ATC-3 markets not having undergone a recall. The graph reports a yearly analysis of the average number of trials in recalled and not recalled ATC-3 groups. The average number of trials is similar in both the ATC-3 markets, having undergone at least a recall (R) and ATC-3 markets not having undergone any recall (NR).

3 Methodology

The main theoretical framework is the same adopted in Acemoglu and Linn (2004). In particular, Acemoglu and Linn (2004) model innovation, the dependent variable of the present model, as being proportional to market size. The measure of innovation is the number of clinical trials in all Phases for the ATC-3 category i. The measure of market size is the sum of products' sales for the i^{th} market. We added to the analysis further potential determinants, time effects, and category effects. The theoretical model returned is the well-known estimation Poisson model:

$$E[N_{it}|\mu_i,\zeta_t,X_{it},M_{it}] = \exp(\beta_1 \cdot \log M_{it} + \beta_2 \cdot X_{it} + \mu_i + \zeta_t) \quad \forall i = 1,\dots N, t = \dots T \quad (3.1)$$

where E is the expectations operator, M_{it} represents endogenous market size, X_{it} captures age (e.g. average age of products in category i weighted for products' size) , diversification and innovation patterns (e.g. scientific production), μ_i are ATC fixed effects and ζ_t time fixed effects. The estimation of (3.1) as written above, however,

would lead to biased estimates for two reasons: first of all, the non-linearity in (3.1) makes it impossible to estimate the fixed effects consistently; secondly, market size is endogenous.

As mentioned in the introduction of the present dissertation, no robust machine learning method has –at the time being – been developed to solve such statistical difficulties.

Thus, in order to deal with both problems, a novel control function (CF) IV approach, described in W. Lin and Wooldridge (2019) has been adopted. With respect to past literature, the present method allows for correcting (i.e., testing and estimating) two potential sources of endogeneity: the one arising from the correlation of the regressors with time-constant, unobserved heterogeneity and the one due to the correlation between covariates and time-varying idiosyncratic errors. Furthermore, it can be easily extended to non-linear scenarios with fixed effects.

Specifically, denoting as κ_{it} the idiosyncratic shock and c_i the individual heterogeneity, the unobserved effects non-linear model allowing for both idiosyncratic endogeneity and heterogeneity endogeneity might look as follows:

$$E[N_{it}|M_{it}, z_{it}, c_i, \kappa_{it}] = c_i \exp(x_{it}\beta_1 + \kappa_{it}) \tag{3.2}$$

where $x_{it} = (M_{it}, z_{it})$. z_{it} would typically include a full set of time effects and M_{it} is the endogenous variable. The exogenous variables, including the vector z_{it}, might correlate with the heterogeneity (i.e., no random effects). There is also a set of excluded exogenous R_{it2} serving as an instrument for the potentially endogenous variable. In the present work, R_{it2} is represented by recalls. W. Lin and Wooldridge (2019) noticed that, without the idiosyncratic endogeneity, an appealing estimator would be a fixed-effects Poisson estimator, which, viewed as a QMLE, would only require a strict exogeneity assumption with respect to the idiosyncratic shocks to ensure consistency. Such an assumption is exploited as a null hypothesis for testing idiosyncratic endogeneity against the alternative of total dependence of the error term of the specification of M_{it} and κ_{it}. The alternative is composed of exploiting the reduced form equation for the endogenous variable

$$M_{it} = z_{it}\Pi + c_{i2} + u_{it2} \quad \forall t = 1, \dots T \tag{3.3}$$

where because the z_{it} is strictly exogenous, it is tested the correlation between κ_{it} and functions of u_{it2}. W. Lin and Wooldridge (2019) developed a simple procedure

allowing to test for idiosyncratic endogeneity and produce consistent estimates also in co-presence of non-linearity, fixed effects and both types of endogeneity. The algorithm follows the steps below:

i) Estimate the reduced form for the endogenous through fixed effects and obtain the fixed effects residuals $\ddot{u}_{it2} = \ddot{M}_{it} - \ddot{z}_{it}\hat{\Pi}$

ii) Use fixed effects Poisson on the mean function

$$E\Big[N_{it}|M_{it}, z_{it}, c_i, \ddot{u}_{it2}\Big] = c_i \exp\Big(x_{it}\beta_1 + \ddot{u}_{it2}\rho\Big)$$

use robust Wald test of $H_0 : \rho = 0$

Step 2 allows estimating the fixed effects in the presence of non-linearity consistently. Yet, fixed effects Poisson enables eliminating ATC-level fixed effects performing a conditional ML consistent estimation. The reader is referred to Cameron and Trivedi (2013) for further details.
A characteristic of Poisson-FE models is that they require the dependent variable to be nonzero for at least one time period. The lower the proportion of zeroes in the dependent variable, the better the model works. The last condition has been fulfilled by dropping those ATC categories not meeting it, constituting approximately 10% of the total ATC-3 in the sample.
We estimated several instances common to literature to check for either delayed effects of trials or the presence of a bias if market size were considered exogenous (or fixed effects omitted).

Throughout, the problem of endogeneity in market size has been exposed as being intrinsic to market size. Hence instrumentation of the endogenous M_{it} is needed. Market size is instrumented through normalized recalls. The normalization is on the number of products present in the market i at time t. Calling m the major recalls, normalized recalls are denoted as follows:

$$\tilde{m} = \frac{m}{\#prod.} \cdot 100.$$

As aforementioned, normalization is necessary to avoid another source of endogeneity. Indeed, ATC markets with more products are more likely to recall by definition. Omitting such control would partly invalidate the estimates. The belief is that markets undergoing major recalls experiment with a sudden negative shock in sales. The

relevance of the instrument is tested in Section 4.

The instrument is not directly related to the dependent variable. The central argumentation that might directly connect normalized recalls to trials is that the lack left by recalls is filled with innovations. Hence, sectors that are more prone to undergo a recall should also be the most innovative ones. In literature, there seems to be contrasting evidence about the topic. Though the argumentation would imply a positive impact of recalls on innovation, the recent events seem to contradict such findings. Indeed, albeit an increasing number of recalls from 2004 to 2015 (see, e.g., Fig.4), the innovation crisis of the pharmaceutical industry is a widely known and recognized phenomenon in literature (see e.g.Pammolli, Magazzini, and Riccaboni, 2011, Price and Nicholson, 2014 among others). It might be argued that the contrasting effects leading to the drop of innovation have overtaken the positive effect of recalls, thus favoring the decreasing pharmaceutical innovation trend. Therefore, the positive effect of recalls on innovation could still be present but hidden. Empirical research has conducted few analyses to explore the relationship between innovation and recalls or withdrawal in general. Fortunately enough, most severe recalls have considerable media coverage, which allowed researchers to collect data on market reaction to such bad events (see, e.g., Pérez-Rodrıguez and Valcarcel, 2012). Authors working on such a stream of literature conclude that the impact of recalls and withdrawals on market innovation has a high variability: some recalls have considerable effects while others have none at all. There seems not to be a systematic way to identify the recalls whose announcement impacted innovation among major recalls. Market reactions depend on not controllable criteria, such as the period during which the recall took place and eventual delays in the FDA's communication of the recall. Generally, however, the market does not systematically overreact to such shocks, invalidating any dependence between recalls and innovation.

To summarize, direct connection sources between innovation and recalls are mainly due to the fixed and time effects. FDA delays cannot be easily controlled. The FDA developed precise guidance and protocols for recall communication and announcement for the period considered in the present Chapter. Hence, delays constitute a minor issue because FDA regulates them. For the sake of completeness, thanks to the FOIA agreement signed, openFDA, and FDA Enforcement report, it has been pos-

sible to verify the happening of delays. The mentioned sources allowed us to access the time gaps between recall initiation, recall classification, and recall termination. The communication of a recall is part of the initiation process. Above all, in case of severe recalls, it must be prompt. The average time between the initiation and the termination for Class I and Class II recalls has been around 23 months. A delay in communication might happen in the first initiation phase. The average time that the initiation phase took for any Class I and Class II recall was four months approximately. For our sample of major recalls, the initiation phase's average time has been approximately 2 to 3 months, in line with prompt communication criteria. This evidence enforces the limited impact of delays on the analysis.

Dropping out unobserved heterogeneity and including time dummies in the primary specification, control for possible direct connections between recalls and innovation. Thus, the mentioned operations ensure only an indirect effect of recalls through sales. Further arguments in favor of the indirect effect of recalls on innovation follow.

In particular, the recalls taken into account are severe recalls of marketed products. The time gap between trial phases and the marketing of a drug usually takes between 8 to 14 years. Such a significant time gap is relevant to guess and understand competitors' possible reactions to a drug recall in the same sector where a firm is operating. We believe that a competitor that underwent a recall in the sector in which both firms operate does not increase or decrease the risk of innovation in the short run. Indeed, marketed products undergo major recalls long after that they are commercialized.

Besides, the lack of sales left on the market by recalling the drug requires an extended period to recover fully. Hence, there is no need to invest in clinical trials to take advantage of such a shortage in the short run. As a further check of the latter conjecture, we build up a time-to-event analysis in Fig.27 of the Appendix. Fig.27 takes into account all types of recall and clearly shows how a recalled product has a truncated life compared to drugs having a normal life cycle.

In particular, Fig.27 displays how the survival rate of drugs that did not undergo a recall is persistently higher than the survival rate of drugs having undergone a recall. Hence, having undergone a recall decreases the "probability of surviving of a drug". Under normal conditions, drugs have a probability greater than 0 to survive more than ten years. However, if a drug underwent a recall, this probability drastically

reduces to almost 0. Notice that the probability that a recalled drug survives two years is still consistent. The median survival time is five years.

Thus, after a recall, the drop in sales is likely to remain unfilled for years. Indeed, had firms found innovative replacements for recalled drug d, which allowed them to recover the shortages left by the recall of d, there would not be any reason to keep selling drug d for years. Recalled products, therefore, leave a long-term lack in terms of sales within the ATC-3 market to which they belong.

A further argument against the coverage of lacks left by recalls through innovative products is that such shortages might be filled by drugs already present in the market, whose trials started before, soon after, or at the same time as the trials leading to the recalled drug. This eventuality is reasonable since, as mentioned, suspended or terminated studies are excluded from the sample, meaning that remaining clinical trials sponsored by concurrent firms are likely to arrive on the market with products belonging to the same therapeutic class. Competition of wholesalers within an ATC might reveal in early stages once it is evident that a firm will develop an innovative cure. The development of alternative drugs is encouraged from the early trial phases when there are still chances to arrive first on the market. Medicines substituting recalled drugs in the same ATC might be developed soon after the recalled medicine in a "first to arrive" competition rather than a "fill the gaps of recalls" logic. The latter may also be because the demand for patented medicines of the type of the recalled drug was likely more consistent when the trial for the recalled drug started. In the eventuality that demand propagates at the recalls' time, either already existing generics or new ones (trials of generics is indeed less time consuming since they only need to ensure bio-comparability) might intervene and fill the gap.

It is worth noticing, in any case, that the potential positive relationship of recalls and innovation exploiting the market lacks passes indirectly through market size. Indeed, the emergence of new trials within a market after a recall depends on the demand that the product in question generated. If a recalled product had no underlying demand, it is reasonable to expect no company to begin a costly trial only to fill the lack left by the recalled product. Therefore, the response of innovation seems to depend on the underlying magnitude of the recalled product's demand, i.e., market size, rather than on the recall itself.

Finally, another possible critique undermining the instrument's validity is that recall of product i might have provoked the recall of trials concerning similar products. This domino effect hangs on the causes of the recall. Indeed, if the recall concerns only the specific product being withdrawn from the market, implications on other companies' products are unlikely. For instance, it is possible that after the recall of the COX-2 inhibitor, Vioxx, due to cardiovascular side effects, all firms having ongoing trials on the same target did suspend or withdraw the trials relating to COX-2 inhibitors. To the best of our knowledge, no effort has been made to explore this possibility in the drugs market. The only work approaching the critique is G. Ball, Macher, and Stern (2018). The authors, however, focus on the medical devices industry, which has different legislation for recalls than the drugs' market. Indeed, a device's recall is a common practice made ordinarily by firms to repair or update a device, which is usually promptly placed back to the market. The way we managed the circumstance is threefold. First, we considered only active trials, thus excluding suspended and withdrawn trials, including those suspended due to other drugs' recall. In a second instance, we also removed trials of companies undergoing a recall. Ultimately, as far as it has been possible to link the reason for the severe recalls [6] we dropped trials adopting a similar active principle. The latter instance happened in a few cases since eliminating suspended and withdrawn trials constitutes already a robust control.

[6] Above all in case of adverse events caused by an active principle adopted in the drug to the scope of a trial

4 Results

The results section is divided into two main subsections. Namely, the impact of recalls on the endogenous market size is first analyzed as measured by total sales of ATC i. The aim is to provide convincing arguments in favor of the relevance of the adopted instrument.
Successively, the results of the impact of the instrumented market size on innovation are presented.

4.1 The impact of recalls on sales

Summary statistics

This Section reports summary statistics for the sample. Tab.8 contains average values and standard deviations (below) of relevant variables for the full sample and two separate sub-samples for observations associated or not to recalls. The Table includes such information at the ATC-3 level and refers to major recalls. Tab.8 embraces all the relevant controls employed for constructing Tab.13.
Tab. 8 displays the overall, between, and within standard deviation for the main controls included sales. The statistics are provided for the total sample, the subset of ATC-3 having undergone at least a recall, and the sub-sample of ATC-3 without recalls. The panel of sales in Tab. 8 displays how, typically, the recalls are found in larger markets than the average. For this reason, recalls have been normalized by the number of products in the ATC market to avoid possible problems of reverse causality with the market size. The normalized recalls have been denoted as $\tilde{recalls}$ in the following paragraphs.
Moreover, as expected, more competitive markets are more prone to recalls, as displayed by the Herfindahl–Hirschman Index (hhi). There is evidence of differences in terms of competition between ATC-3 groups. The Appendix provides further insights into which type of firms and products generally undergo a recall. Specifically, in Appendix 3 we evidence a general tendency of recalls to be located in big established firms and to regard relatively older products than the average age.
On the contrary, with respect to firm and product levels, recalls are located in more

Table 8: Summary statistics at the ATC-3 level for the full sample, the subset of the ATC-3 having undergone a recall in the period considered, and the subset not having undergone a recall. Database at the ATC-3 level is balanced.

Variable		ATC-3			Description
		Full Sample	Subs. recalls	Subs. no recalls	
Sales (log)	Overall mean	19.405	20.794	19.054	Log of sales at ATC-3 level.
	Overall Std. Dev.	2.233	1.475	2.256	
	Between Std. Dev.	2.152	1.441	2.163	
	Within Std. Dev.	.614	.378	.661	
Outflow rate ($\frac{K_{t+1}}{P_{-1}}$)	Overall mean	.086	.056	.093	It is defined as the number of lost products in an ATC-3 (K_{t+1} in regressions) over the total number of products in $t-1$ (P_{-1} in regressions).
	Overall Std. Dev.	.257	.064	.285	
	Between Std. Dev.	.109	.036	.120	
	Within Std. Dev.	.233	.053	.259	
Avg. age of firms within ATC	Overall mean	35.907	33.275	36.573	It is the average age of the firms competing within an ATC-3. The foundation year of the firms was present in the data.
	Overall Std. Dev.	7.631	5.420	7.960	
	Between Std. Dev.	6.767	4.452	7.093	
	Within Std. Dev.	3.556	3.160	3.650	
Herfindahl–Hirschman Index (hhi)	Overall mean	.431	.268	.434	The hhi measures the competition within a market. It can range from 0 to 1.0, moving from a huge number of very small firms to a single monopolistic producer.
	Overall Std. Dev.	.260	.159	.262	
	Between Std. Dev.	.236	.166	.240	
	Within Std. Dev.	.110	.020	.115	
Share generics by ATC	Overall mean	.746	.725	.752	It represents the percentage of generic products, among all products sold in an ATC-3 market
	Overall Std. Dev.	.255	.214	.264	
	Between Std. Dev.	.238	.210	.245	
	Within Std. Dev.	.092	.052	.099	
Avg. age prod. by ATC	Overall mean	13.159	12.043	13.441	It represents the average age of product within an ATC-3. The age of a product is based on the foundation year of the firm that produced it.
	Overall Std. Dev.	5.333	3.763	5.627	
	Between Std. Dev.	4.909	3.568	5.164	
	Within Std. Dev.	2.109	1.304	2.268	
Scientific knowledge within ATC	Overall mean	6.327	6.787	6.211	The number of papers and scientific publications for an ATC-3 present in PubMed and other sources.
	Overall Std. Dev.	1.718	1.623	1.724	
	Between Std. Dev.	1.705	1.627	1.709	
	Within Std. Dev.	.242	.177	.256	
Number of firms within ATC	Overall mean	21.054	32.802	18.082	Number of firms trading within an ATC-3
	Overall Std. Dev.	20.954	23.442	19.175	
	Between Std. Dev.	20.604	23.069	18.876	
	Within Std. Dev.	4.050	5.367	3.645	

dynamic ATCs, where recalled drugs were pioneering in the past. Fixed effects technique accounts for time-invariant characteristics of ATCs.
The recalls intervene in firms with a high share of generics (see Appendix 3). This finding might result from a less stringent policy for generic drugs' approvals than branded ones. Growing concern for generic safety is, in fact, a well-known problem in literature (see, e.g., Gallelli et al., 2013).
Besides, in ATC markets, the outflow rate presents a within variance higher than the between variance. The latter means no difference between ATC-3 groups concerning the outflow rate. As opposed to the firm level, this inversion is expected. Indeed, while strategic policies of product placement might occur in firms, this is not the case for ATC aggregation, where market laws apply. Thus, on average, even two utterly different ATC markets would display similar outflow rates following only a demand-supply logic.
Two other variables seem to be related to recalls at the ATC-3 level, i.e., scientific knowledge within an ATC and the number of firms trading within an ATC. Specifically, the recalls happen in ATC markets where, on average, trade more firms and scientific knowledge is more advanced than other markets.
To summarize, the recalls regard relatively old drugs produced in big established firms. The major recalls occur in relatively dynamic markets whereby, on average, many younger firms operate, trading relatively young products. A possible reason the markets having the described characteristics undergo more easily recalls is that they are precisely the markets monitored by the legislator with special attention.

Analysis of the determinants of drug recalls

This section reports the first-stage results. A Fixed-Effects estimation method is employed. Tab.9 shows the estimates of the first stage at the ATC-3 level. As detailed below, a different level ATC-Firm has been added to test for compensations within ATCs inside firms. For consistency with the best model, the sample was truncated in 2013 also for the first stage. Outcomes with a not-truncated sample display very similar results (see Appendix 3). The F-statistic amounts to 14.32.
The standard errors included in the Tables of the present Chapter are all robust and clustered at the ATC-3 level of aggregation.

Table 9: First stage results at different levels.
ATC-3 aggregation represents the main specification.

	(ATC-Firm Aggregation) Log sales	**(ATC-3 Aggregation)** Log sales
$\tilde{recalls}$	−0.0053	−0.0283***
	(0.0033)	(0.0056)
$\tilde{recalls}_{t-1}$	−0.0226**	−0.0267***
	(0.0083)	(0.0070)
$\frac{K_{t+1}}{P_{-1}}$		0.1932**
		(0.0628)
average age firm		0.1576
		(0.0921)
average age firm^2		−0.0020
		(0.0013)
hhi		1.2405***
		(0.2590)
share generics in ATC		−0.1895
		(0.3373)
papers		−0.0260
		(0.0507)
# firms		0.0077
		(0.0071)
Year Dummies	Yes	Yes
Obs.	48915	1664
Groups	8634	208

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered (at the ATC-3 level) standard errors are in parentheses. First-stage results are shown in this Table. (1) fits an F.E. model at the ATC-Firm level, i.e., ATC lines of productions within firms. This level is introduced to check the possibility of compensations between sales of products belonging to the same ATC (excluded due to the significance of the coefficient of recalls witnessing a drop after a recall) (2) fits an F.E. model at the ATC-3 level. $\tilde{recalls}$ represent recalls normalized. At the ATC level, recalls are respectively normalized for the number of products within an ATC.

We found a significant and negative impact of recalls on the logarithm of sales at the market level. In the Appendix, it is shown that sales of firms undergoing a recall are unaffected. At the same time, the lines of production of medicines belonging to the same ATC-3 encounter a drop in sales due to recalls (ATC-Firm Aggregation in Tab.25 and 9). This evidence excludes the possibility of compensations between sales of products belonging to the same ATC inside a firm. Therefore, the negative effect of recalls at the market level is enforced, whose lacks are not filled by the same firms with other medicines of the same ATC-3.

The second column of Tab.9 represents the first stage of the principle analysis. As illustrated, the effect of recalls at the ATC-3 level is powerful and significant for current recalls and delayed ones. After having performed a sufficient amount of bootstrap repetitions, we found that the t-statistic is invariant to whether we use recalls or lag recalls to obtain it [7]. This finding corresponds to a Sargan-Hansen test for over-identification in our contest, implying the absence of over-identifying restrictions (W. Lin and Wooldridge, 2019). We believe that the key reason for the strength of the result relies on the level of aggregation. While firms with high-quality managements and inclined to risk can promptly make up for severe recalls, the latter take ATC-3 markets unaware. Competitors could not anticipate severe recalls against firms producing in the same ATC as theirs, which can be detected only at the market level.

The absence of compensations at the market level has been further tested. In particular, we analyzed the effect of recalls on aggregated sales once the firms' sales having undergone a recall are removed from the sample. The drop in sales seems to disappear once firms having undergone a recall are excluded (see Fig. 31 in Appendix).

The fall of sales observed at the ATC-3 level becomes evident not only from the estimates in Tab.9 but also from the study of abnormal values in Section 4.1

Finally, it might be argued that since recalled products are the most innovative ones, no direct substitute is present in the same market. However, our dataset provided the generic name of products (both recalled and not recalled) and the active principle of medicines. It has thus been possible to detect an average of 10 products within the market exploiting the same active principle as the recalled products. Hence, it has also validated the hypothesis that the lacks left by the recalls might be filled with

[7] t-stat is obtained after 30000 repetitions and amounts to 2.438

products already present on the market and that recalled products are not necessarily the most innovative ones having no substitutes.

Analysis of Abnormal Values

This Section reports estimates of the influence of drugs' recalls on sales. The effect of recalls is defined by taking a reference value of the given economic indicator as it would be observed under "normal" dynamics of economic conditions; this is called the "potential" value. We hence define the Abnormal Value (AV) of the indicator y associated with the unit i in time t as the difference between the observed and the potential value Thirumalai and Sinha, 2011:

$$AV_{it} = y_{it} - E\left(y_{it}\right), \tag{3.4}$$

The potential value $E(y_{it})$ is estimated by running a Fixed-Effects regression on the following model:

$$y_{it} = \alpha + \beta y_{st} + \gamma X_{it} + \mu_i + \lambda_t + u_{it}, \tag{3.5}$$

where y_{st} is the aggregated value of y in year t at the sector level. The control variables (X) and the year dummies are included as regressors. After obtaining estimates of AV_{it} for all i and t, referred to as $\widehat{AV}_{it}$, the time dimension is re-scaled. The latter exercise allows the time variable to be centered on the year when the recall is issued. The change occurs for all units experiencing a recall in the time frame considered. Only these observations are kept in the sample. The market-level Abnormal Value $\overline{AV}_t$ associated to recalls is then computed as the simple average of $\widehat{AV}_{it}$ for any $t \in \{-(T-1), ..., (T-1)\}$, as follows:

$$\overline{AV}_t = \sum_{i=1}^{N_t} \widehat{AV}_{it}, \tag{3.6}$$

where N_t is the number of units with available data in t among those experiencing one recall. Confidence intervals for $\overline{AV}_t$ are constructed calculating the variance of $\widehat{AV}_{it}$ as follows:

$$Var\left(\overline{AV}_t\right) = \frac{\sum_{i=1}^{N_t} Var\left(\widehat{AV}_{it}\right)}{N_t^2}, \tag{3.7}$$

where $Var(\widehat{AV}_{it})$ is the variance of the forecast error derived from estimation of Equation 3.5. The focus of the analysis is on the growth rate of sales volumes. The exercise is replicated for three classifications of recalls (standard recall definition, major recalls, type of recall) and three levels of analysis: product, firm, and sector level. The main text reports only the analysis at the ATC-3 level as it is the level at which the first and second stages are conducted. Abnormal values at the firm and product level can be found in Appendix.

Note that in the model for the sector level, y_{st} is replaced with y_{mt} in Equation (3.5), that is the value at the entire market level.

Fig. 5 reports estimates of the effects of recalls on the AV of sales growth.

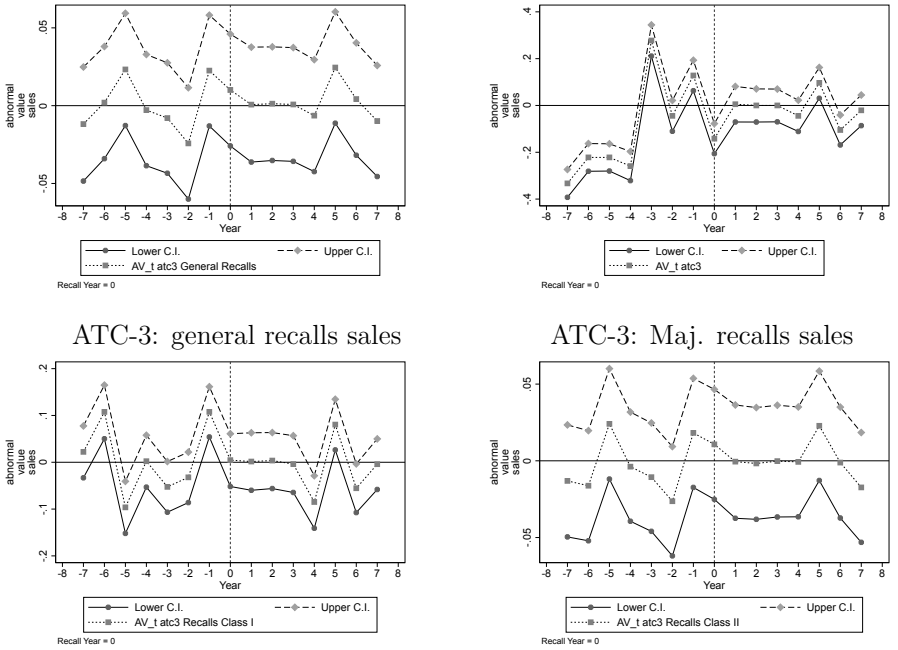


Figure 7: Abnormal values at ATC-3 level of aggregation. Years are normalized. Year 0 represents the year of the recall. The four scenarios include the path of sales before and after the recall year, using four different definitions of recalls: major recalls, Class I recalls, general recalls, and Class II recalls. As it is evident from the diagrams, sales drop at recall year for every type of recall. Major recalls presenting a more pronounced drop. Moreover, using major recalls, the lowest error bound is reached.

Fig.7 exhibits abnormal values for ATC-3 level. Confidence intervals are constructed at the 95% level. As it is evident from Fig.7, after the initial drop at the year of recall, sales soon recover one or two years after year 0 (see major recalls).

The latter observation classifies the instrument employed in our work as a short-run effect. This distinguishes the effect of our instrument from the long-run effect that demographic shocks produce in the work of Acemoglu and Linn (2004).
The analysis of abnormal values confirms what was found in previous paragraphs. Especially a considerable impact of recalls on sales in the year of the recall. The error bound is lower for the ATC-3 level with major recalls, thus enforcing the expectation of a drop in sales at the recall time.

4.2 Relation between innovation and market size

In this section we report the results concerning the relationship between market size, M_{it} and innovation, N_{it}. Since the data at our disposal are already converted into dollars of 2015 using Consumer Price Index (CPI), market size is measured directly as the sum of sales over ATC market i at time t. Innovation is measured with the number of activated trials in ATC i at time t. The time window ranges from 2004 to 2013. The samples' last two years (2014, 2015) have been cut away since very few trials have been conducted in such a period. Including 2014 and 2015 may have led to biases in the procedure, which exploits Poisson estimates. Indeed, the latter method does not tolerate a value of 0 for the dependent in most observations.
The panel is strongly balanced as required by the procedure. Each year has data for 208 therapeutic classes.
The best model is estimated by Eq.(3.1).
We introduced several regressors. These comprise supply-side determinants, technological opportunities, and age determinants. We draw some controls directly from the literature, comprising knowledge stock (see, e.g., Cerda, 2007, Acemoglu and Linn, 2004 among others) as measured by the number of papers referred to ATC category i. PubMed database has been consulted. Specifically, we collected the number of scientific works for a given ATC-3 in a given year through Mesh Terms. According to NIH, MeSH terms are official words or phrases selected to represent particular biomedical concepts. When labeling an article, indexers select terms only from the official MeSH list, never other spellings or variations. For deciding whether a paper referred or not to a specific ATC class, it has been first associated a Mesh Term to ATC category i primarily exploiting the official synthetic description of ATC. If the latter did not

produce any result or did not match evidence from the literature, a double-check was made using level 3 indications as Mesh terms.[8]. NCBI Mesh database allowed us to customize the searches. Since the number of papers showed an upward trend, the variable has been detrended through first differentiating its logarithm.

Another critical control drawn from literature is the share of generics. As noted in Dubois et al. (2015), ease of entry and substantial financial incentives to use generics will reduce the expected profitability of the innovation. Hence it is vital to detect the degree of penetration of generic medicines within markets to discourage firms from undertaking innovation.

Besides, as emphasized both in Acemoglu and Linn (2004), and Dubois et al. (2015), a further source of declining margins of innovation is represented by the increasing number of young entrants within an ATC market. Pharmaceutical competition, in general, might undermine innovation productivity. It is, thus, imperative to measure and control for competition.

Apart from Acemoglu and Linn (2004), empirical literature does not model explicitly competition (see Dubois et al., 2015). In the present work, we constructed two measures to control pharmaceutical competition. The first is the Herfindahl index (*hhi* hereafter), which measures firms' size concerning the market. It is usually employed as an indicator of competition among firms within an ATC-3 market. The Herfindahl index's major benefit compared to other measures such as the concentration ratio is that it gives larger firms more weight. The index can range from 0 to 1.0, moving from many tiny firms to a single monopolistic producer. The second measure controlling competition is the average age of firms within a market. It controls other aspects of competition compared to *hhi*. While *hhi* measures the "degree of monopoly" within an ATC, it cannot clarify the firms populating the market. However, the average age of firms mainly catches the presence of small biotechnology firms in the market. Such firms are known on one side to compete for innovation and on the other to have less financial resources in contrast with established companies (see, e.g., B. H. Hall and

[8]For instance, category C6B is described as "PULMONARY ARTERIAL HYPERTENSION (PAH) PRODUCTS." Due to the name's length and possible different abbreviations employed in the Mesh Terms list, Mesh Terms have been searched by looking at different specifications of the description such as "PAH PRODUCTS," "PULMONARY ARTERIAL HYPERTENSION PRODUCTS." If the latter did not produce any result or the results were not in line with the findings in the literature, then the Mesh indication at level 3, "PAH" was selected

Rosenberg, 2010 among others). Since margins decline with the number of young entrants, we expect a negative sign of firms' average age.

Tab.10 presents the main results of the analysis. It is technically the second stage of the procedure described in the methodological section. Precisely, calling z_{it2} the excluded instruments ($\text{recalls}_{it}, \text{recalls}_{it-1}$), the first stage estimation computes the residuals, $\ddot{u}_{it2}$, of a linear fixed-effect model whose dependent is market size. The second stage incorporates the residuals and estimates a fixed effect Poisson model. Please refer to steps 1. and 2. in the methodological section.

Differently from literature, in the present work, it is not necessary to construct M_{it} based on demographic shifts since the innovative instrument, recalls, already purges market size from endogeneity. In the following, M_{it} is simply the logarithm of collapsed sales at ATC-3 level, i.e., the product of the number of purchased drugs expressed in standard units to ensure comparability with their price.

Notice that a critical assumption of the model is that excluded exogenous, R_{it} appearing within z_{it}, do not explicitly appear in the equation of Trials. For the more refined aggregation level at our disposal, ATC-3, it is plausible to assume that the average elasticity is the same across categories.

Table 10: Impact of market size on innovation. Col.(1) employs a simple Poisson model not considering fixed effects. Col.(2) is the main specification (fixed effect Poisson). Col.(3) and Col.(4) add the lag of the dependent. Col.(5) eliminates all the controls

	(1) Trials	**(2)** Trials	**(3)** Trials	**(4)** *log Trials*	**(5)** Trials
$trials_{t-1}$			−0.00741	0.0732*	
			(−1.45)	(0.0335)	
Log sales	0.1378***	0.6362**	0.802**	0.1176***	0.8229**
	(0.0060)	(0.2149)	(3.01)	(0.0153)	(0.3174)
residuals		−0.8018***	−0.862**		−0.9711**
		(0.2157)	(−3.21)		(0.3177)
$\frac{K_{t+1}}{P_{-1}}$	−0.5378***	−0.0926	−0.484***	−0.0504	
	(0.0847)	(0.0909)	(−3.30)	(0.0914)	
average age firm	0.2890***	−0.1332***	−0.106*	0.0634**	
	(0.0139)	(0.0377)	(−2.66)	(0.0214)	
average age firm^2	−0.0038***	0.0021***	0.00178***	−0.0008**	
	(0.0002)	(0.0005)	(3.31)	(0.0003)	
hhi	0.2245***	−0.3199	−0.145	0.1106	
	(0.0446)	(0.2903)	(−0.40)	(0.1153)	
share generics in ATC	−0.5571***	−0.3168**	−0.898***	−0.2036	
	(0.0404)	(0.1124)	(−6.29)	(0.1068)	
average age product	−0.0658***	−0.0592**	−0.0928**	−0.0564***	
	(0.0061).	(0.0190)	(−2.88)	(0.0143)	
average age product^2	0.0010***	0.0011	0.0100	0.0010*	
	(0.0002)	(0.0010)	(1.64)	(0.0004)	
papers	0.5608***	0.1558*	0.101	−0.0443	
	(0.0728)	(0.0750)	(1.22)	(0.1477)	
papers^2	−0.6672***	−0.0067	−0.0929	−0.0083	
	(0.1056)	(0.0838)	(−1.17)	(0.0883)	
# firms	0.0090***	0.0008	−0.0032	0.0048**	
	(0.0007)	(0.0035)	(−0.70)	(0.0016)	
Year Dummies	Yes	Yes	Yes	Yes	Yes
Obs.	1664	1664	1664	1664	1872
Groups	208	208	208	208	208

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered (at the ATC-3 level) standard errors are in parentheses. (1) fits a simple Poisson with exogenous sales. (2) represents the main specification. The dependent variable is the count of active trials in ATC i at time t. The time interval is ten years. The technique adopted for the estimation is Wooldridge [2019]. Please refer to Section. 3. (3) count model with lagged dependent among regressors following Acemoglu and Linn, 2004 (4) linear model with lagged dependent among regressors and exogenous size. Dependent is linearized. Both the presence of non-linearities and endogeneity are ignored. (5) best model without controls

Column (1) presents a simple Poisson model with exogenous market size exploring whether market size's positive effect is robust in the absence of fixed effects and endogeneity controls. Column (2) is our main specification, i.e., a fixed effect Poisson controlling for market size's endogeneity.

The coefficients of interest in Tab.10 are Log sales and residuals. The former represents the market size, and the latter measuring endogeneity of market size. Specifically, a significant coefficient of residuals means a correlation between the error term (see the specification in Tab.10, i.e., second stage regression) and functions of the error of the model of the market size (first stage). In other words, residuals control for co-movements of sales and unobservables related to the number of trials. Market size is hence "purged" from the alleged endogenous part. Endogeneity is tested with a Wald test on residuals' coefficient ρ. If ρ is significantly different from zero, endogeneity is present. This latter instance occurs in our model as expected (Column (2)). In particular, fully robust standard errors detect a strong idiosyncratic endogeneity. The exploitation of Fixed Effects methodologies allows the unobserved heterogeneity to be correlated with all explanatory variables and the excluded exogenous recalls. The evidence is that even after allowing the market size to be correlated with the ATC heterogeneity, market size is not exogenous to idiosyncratic shocks.

The coefficient of market size is positive and significant in line with past works. According to our estimate, a 10% increase in market size leads to an increase of almost 6.3 % of active trials. It turns out that also the magnitude conforms with literature. Indeed, previous research generally finds elasticities to be approximately 0.5 consistently with our estimates.

Recent literature speculated on the possibility that, though clinical trials might respond elastically to market size, the proportion of them resulting in effective innovation might decline (see e.g.Dubois et al., 2015 among others). Hence authors might have overestimated the effect of market size on clinical trials since the latter should be computed only on the trials that effectively brought innovation. This chapter exploits active trials as a dependent, which partially solves the issue. We believe that active trials constitute the subset of promising trials in terms of innovative contribution. The estimated higher effect than the literature that adopts NMEs or NCEs as a dependent is well explained by the substantial costs for developing new pharmaceutical

entities. Drug development is, in fact, quite expensive, the cost ranging between $800 Million to $2.5 Billion (see, e.g., MedChem (2012) [9]). Undertaking clinical trials is, instead, sensibly cheaper, amounting to an average of $20 Million to $40 Million (see Martin et al., 2017 as well as John Hopkins Bloomberg Health School, 2018). Thus it is reasonable to suppose that, ceteris paribus, a 10% increase in market size stimulates more trials than NMEs or NCEs on average. Exceptions are still present (see Acemoglu and Linn, 2004, Duggan and Scott Morton, 2010, who estimated an higher elasticity than the one of the present work).

The coefficient of the average age of firms and its square is in line with past observations (see, e.g., Huergo and Jaumandreu, 2004 and Balasubramanian and J. Lee, 2008 for specific studies on the topic). The effect evidences how the oldest firms tend to introduce less innovation than entrants in their early years. However, firms above intermediate ages *appear almost as active in process innovations as entering firms and even more in product innovations* (Huergo and Jaumandreu, 2004).

Moreover, innovation decreases with the share of generics within a market. Thus, the effect theorized in Dubois et al., 2015 of decreasing margins of innovation proportionally to the entrance of generics reveals to be correct (see also Lanjouw, 2005).

In line with Acemoglu and Linn, 2004 and Rake, 2017, technological advancements as measured by detrended papers are positively related to innovation. It is reasonable to suppose that more trials emerge in markets where scientific research is prolific.

The discrepancies in the magnitude of the coefficients between the main specification (Column (2)) and Column (1) of Tab.10 can be explained in several ways. In Column (1) of Tab.10 correlation over time of units is not controlled. So it is assumed that units are independent over the cross-sectional dimension and over time dimension, which is quite a strong constriction in a longitudinal setting. The assumption means that the same individual (market) observed at two different times, t_0 and t_1, is considered independent from herself. In other words, individual (market) i at time t_0 is another individual (market) than individual (market) i at time t_1. The main implication of such presumption is that unobserved time-independent heterogeneities of individuals do not affect other individuals. However, we know that the same individual observed at two different times is considered "two distinct individuals." Thus,

[9] n.d.

in the model of Column (1), it is ultimately assumed that unobserved shocks of an individual (market) i at time t do not influence individual (market) i at time $t+k$. In other words, we are mixing between and within individual effects. Between effects are obtained once the time component is averaged out from the variables. Between-effect settings exploit differences between units, which in our case are independent by definition (we take ATC-3 markets, see previous Sections), not taking into account time variations. Therefore, the market size variance (time-demeaned) will be higher in a between-effect setting since it considers the average market size difference between independent ATC-3 markets. Furthermore, given the opposite time trends of trials and market size (see Fig. 3) in a between-effect setting, the between effects of market size on innovation will be deflated. Indeed, the innovation trend decreases from a specific time on while the market size trend increases. However, since time variations are not controlled in a between-effect setting, the inverse proportionality of market size and innovation emerges. Mixing between and within individual effects will, hence, result in an overall lower coefficient of Column (1) compared to Column (2).
Ultimately, the downwardly biased coefficient of Column (1) suggests that the unobserved heterogeneity is negatively correlated to trials.
To provide an example, consider the possibility that an ATC experienced a sizeable positive shock (more trials) in 2010. For some reason, the mentioned shock is not modeled nor measured. All else being equal, the apparent fixed effect for that ATC in the period 2004-2013 will appear to be higher. However, from the literature, we know that the more the products available for treating a particular clinical condition, the lower the margins on each product (see Bresnahan and Reiss, 1991 among others). The unobserved positive shock for ATC i^{th}, therefore, would lower the margins of all competitor products in the same market, pushing down the sales for the same market. This negative correlation between the market size regressor and the error term deflates the estimate for market size. Vice versa, in Column (2), time dependency is controlled, and deflation is eliminated. Therefore, the coefficient of market size results is higher than in Column (1). Column (1) does not control the reverse causality of market size on innovation. Not considering the reverse causality of market size contributes to upward biasing the market size's coefficient (see, e.g., Acemoglu and Linn, 2004). There are, therefore, in Column (1), two contrasting effects: the up-

ward effect due to the reverse causality endogeneity and the downward bias given by the unobserved heterogeneity endogeneity. The two effects do not compensate, and negative heterogeneity bias prevails over reverse causality endogeneity bias.

Robustness checks Col.(3)-(5) of Tab.10 investigate the robustness of the effect of market size on innovation. Three additional models are added to the preferred specification. Precisely, Column (3) reproduces the exercise of Acemoglu and Linn (2004) to control for possibly varying over time technological flows (see below) by adding lagged trials among the regressors. Since the estimating equation in Column (3) is nonlinear, we perform this instrumentation strategy by adding the residuals of the first stage. Column (4) is the same as Column (3), where the dependent variable is log linearized, and residuals are ignored. Column (4) ignores both the presence of non-linearities and endogeneity.

Adding lags of the dependent variable is a valuable exercise. Indeed, following Acemoglu and Linn (2004), the primary threat to the identification strategy of innovation is represented by changes in the flow rate of innovation for every dollar spent for research on a drug (permanent differences in innovation are already dropped through the ATC fixed effects). Differences in the flow rate of innovation suggest that technological progress is scientifically more difficult in some lines than others. The parameter denoting innovation flow is part of the theoretical specification of innovation drawn from Acemoglu and Linn (2004). Following Acemoglu and Linn (2004), if the flow rate of innovation varies over time, it is also likely to be serially correlated. Adding lag of log innovation to the preferred specification is a simple way to check the importance of these concerns. Lagged market size is instrumented with its lags through a system GMM one-step procedure. The p-value of the Hansen test of overidentification of model in Column (4) is 0.175, falling mainly between the tolerance levels of 0.1 and 0.25 indicated in Roodman (2009). The latter control ensures the validity of the system GMM instruments employed. The Arellano-Bond test is investigated in Tab.11.

	z-score	p-value
Arellano-Bond test for AR(1) in first differences:	z = -10.47	Pr > z = 0.000
Arellano-Bond test for AR(2) in first differences:	z = 0.88	Pr > z = 0.377
Arellano-Bond test for AR(3) in first differences:	z = -1.46	Pr > z = 0.145
Arellano-Bond test for AR(4) in first differences:	z = 0.33	Pr > z = 0.740

Table 11: Arellano-Bond test for autocorrelation of first differenced residuals of GMM

When the idiosyncratic errors are independently and identically distributed (i.i.d.), the first-differenced errors are first-order serially correlated. So, as expected, the output above presents strong evidence against the null hypothesis of zero autocorrelation in the first-differenced errors at order 1. Yet, as suggested in Roodman (2009), *"in the context of an Arellano-Bond GMM regression, which is run on first differences, AR(1) is to be expected, and therefore the Arellano-Bond AR(1) test result is usually ignored in that context"*. The output above presents, moreover, no significant evidence of serial correlation in the first-differenced errors at orders 2, 3, and 4.
Market size is considered exogenous in Column (4), though fixed effects are controlled. The model in Column (4) is linear. In order to ensure comparability among models, trials have been transformed to a logarithmic scale. Column (4) is, in other words, an essential control since, though controlling for fixed effects, it ignores the presence of potential non-linearity (misspecification) and endogeneity, proposing the hypothesis of serial correlation.
Finally, Column (5) presents the model without further control as estimated by the preferred specification's control function approach. The idea beyond Column (5) is to check whether not controlling for regressors compromises the main specification estimates.

The outcomes of Col.(3)-(5) of Tab.10 confirm the estimates of the main specification for what concerns the positive effect of market size on innovation.

Columns (3)-(5) in Tab.10 all display a positive effect of market size on innovation. Expressly, Column (3) confirms the results of Acemoglu and Linn (2004) finding no evidence of serial autocorrelation. In particular, the coefficient of lag trials is negative and non-significant as in Acemoglu and Linn (2004). Possible explanations are already in Acemoglu and Linn (2004) and are, therefore, not discussed in the present work. In Column (4) of Tab.10, the positive coefficient of lagged trials is significant at the 5% tolerance level. This evidence is almost in line with Acemoglu and Linn (2004) when no instrumentation is performed. [10] Under this scenario, the lagged dependent's coefficient turned out to be positive and not significant also in Acemoglu and Linn (2004). Market size is strongly and positively related to innovation, with a coefficient having the lowest magnitude of the specifications analyzed. Indeed, some of the variability might be caught by lagged dependent. Moreover, possible misspecification bias might intervene due to the not correction of nonlinearity.
Notice that the effect of the market size in Column (4) of Tab.10 displays similarities to Col. (1) of the same table, which does not control for endogeneity. Furthermore, market size is more prominent in the models correcting for endogeneity. Therefore, in general, the lack of control for temporal dependence may matter very little for estimation, as it is also consistent with the fact that the autocorrelation coefficient is very weak. Otherwise, indeed, also the coefficient of size in Columns (1) and (4) of Tab.10, whose only dissimilarity relies on the control for temporal dependencies (Col.(4)), would have sensibly differed. Hence, it is reasonable to suppose that the lower magnitude of the coefficient of size in both Columns (1) of Tab.10 and Column (4) of Tab.10 is primarily a consequence of considering market size as exogenous. It is possible to provide further checks by controlling for possible overidentifying conditions of the instrumented lagged dependent variable. To do so, Tab.12 column (1) reports the two-step robust GMM estimates of Column (3) of Tab.10, which, instead, performed a system one-step GMM.

[10] different from Column (1) where residuals of first-stage are included

Table 12: Coefficients of market size and lag dependent when a two-step GMM is employed. Col.(2) includes suspended and withdrawn trials in the dependent

	(1) *log Trials*	(2) *log Trials*
$trials_{t-1}$	0.0592	0.380*
	(0.0426)	(0.189)
Log sales	0.1208***	−0.533**
	(0.159)	(0.179)
Year Dummies	Yes	Yes
Obs.	1664	1664
Groups	208	208

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered (at the ATC-3 level) standard errors are in parentheses. (1) is the two-step GMM version of Column (3) Tab.10, which performed a system one-step GMM. The Table comprises only the critical coefficients. (2) is equal to (1), where suspended and withdrawn trials are included in the dependent. Both equations are linearized to enable a simple comparison with Column (3) Tab.10. The same results apply if the count of trials is employed as a dependent variable. (see F., 2021)

The coefficient of the market size in Tab.12 is higher compared to the one in Column (3) of Tab. 10. Furthermore, the lagged dependent variable is not significant in line with Acemoglu and Linn (2004). The same applies if the count of trials is employed as a dependent (see Appendix).

The sign of the estimates of market size does not change with respect to the preferred model.

A final robustness check has been made by including all the trials, i.e., active ones and suspended and withdrawn. The number of classes employed is the same, though the number of trials increased by 0.57% on the total. This is performed in Tab.12 column (2). As displayed, our estimation does not confirm the hypothesis in Dubois et al. (2015) showing a lower coefficient instead both in terms of magnitude and significance level. The hypothesis is that non-active trials, which are less responsive to market size, bias the estimates leading them toward randomness. For instance, firms with all suspended trials are unaffected by price regulations that reduce the price of treatments by governments. Simultaneously, increases in market size could be less effective on such companies, which already have sunk costs due to inactive trials. The presence of endogeneity is confirmed.

Further robustness checks have been performed by changing the market size's proxy

to align to Acemoglu and Linn (2004), moving to another database to collect sales data (Evaluate sales are employed), and employing all the recalls at our disposal to instrument market size. In particular, Tab.13 shows the outcomes of the analysis adopting Class II and Class I recalls as an instrument for market size. Tab.14 measures market size through the number of patients within an ATC-3 in accordance with Acemoglu and Linn (2004).

Since the number of patients is highly correlated with sales and it is employed as a natural alternative to sales, we adopted recalls as an instrument for the number of patients.

Furthermore, the F-test amounts to 12 for the analysis with Evaluate and 4 for the analysis with the number of patients.

Table 13: Col.(1) and Col.(2) represent first and second stage results using all the recalls at our disposal. Data are aggregated at the ATC-3 level.Impact of market size on innovation using Evaluate database (Col.(3)) and number of patients (Col.(4)) as proxy of market size

	(First-stage all recalls)	**(Second-stage all recalls)**	**(First-stage Evaluate)**	**(Second-stage Evaluate)**
	Log sales	Trials	Log sales	Trials
$\tilde{recalls}$	0.00199		-0.260^{***}	
	(0.64)		(0.0059)	
$\tilde{recalls}_{t-1}$	-0.0223^{***}		-0.0180	
	(-3.34)		(0.0105)	
Log sales		0.580^{*}		0.710^{**}
		(2.02)		(0.275)
Residuals		-0.739^{*}		-0.724^{***}
		(-2.13)		(0.275)
$\frac{K_{t+1}}{P_{-1}}$	0.191^{*}		-0.173	-0.147
	(3.07)		(0.154)	(0.276)
average age firm	0.153	-0.129^{*}	0.0470	0.224^{***}
	(1.67)	(-2.28)	(0.0678)	(0.0045)
average age firm^2	-0.00195	0.00207^{**}	-0.0004	-0.0814^{***}
	(-1.53)	(2.87)	(0.0008)	(0.0316)
hhi	1.238^{***}	-0.250^{***}	1.721^{***}	-0.60
	(4.79)	(-0.56)	(0.419)	(0.495)
share generics in ATC	-0.178	-0.326^{**}	-0.372	0.00640
.	(-0.53)	(-2.69)	(0.328)	(0.157)
average age prod.	0.0539	-0.0559^{*}	-0.0330	-0.0732^{**}
	(0.92)	(-2.33)	(0.0504)	(0.0225)
average age prod.^{2}	-0.0037	0.0008	0.00138	0.0009
	(-1.72)	(0.63)	(0.00206)	(0.0009)
papers	-0.0267	0.153^{*}	-0.139	0.265^{**}
	(-0.52)	(0.0507)	(0.0803)	(0.0917)
# firms	0.00729	0.00103	-0.0106	0.0224^{***}
	(1.04)	(0.25)	(0.0095)	(0.0045)
Year Dummies	Yes	Yes	Yes	Yes
Obs.	1664	1664	1136	1056
Groups	208	208	142	132

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered (at the ATC-3 level) standard errors are in parentheses. The Table shows the results when all the recalls at our disposal are employed. First-stage results are shown Col. (1) while second stage results are in Col. (2). The level of aggregation is the ATC-3 one and the estimation method is the same as the one adopted in the main analysis.

Tab. 13 reports the first and second stages of employing Class I and Class II recalls as an instrument for market size in the first two columns. The remaining columns are devoted to the results obtained using Evaluate database to collect market sales. The results of the primary analysis are confirmed in both exercises.

Employing all recalls decreases both the magnitude and the significance of the coefficients of sales. Moreover, only the lag of recalls is a good instrument at the market level. These two effects are expected since minor recalls may attenuate the drop in

sales consequent to a recall. Indeed, within Class II, temporary recalls (e.g., recalls due to a labeling error) are also comprehended, which may not be unexpected to the firm (most of them are voluntary). For this reason, they might not be taken into account by the company's management. Losses in terms of sales are, therefore, well-compensated. Furthermore, minor recalls are not publicized and cannot damage the image of the company or the market in which they happen.

Hence, adding minor recalls overtakes the strong and negative impact of current recalls and, consequently, affects the market size estimates in the second stage. Since, however, Class II recalls often regard minor but persistent issues[11], a cumulative effect intervenes, and lagged recalls remain an excellent instrument.

The outcomes of the primary analysis remain robust when data on sales are collected from a different database.

Tab.14 reports the second stage results of the analysis with the number of patients as a measure for market size. First stage results are in the Appendix 3.

[11] minor recalls often pertain to the manufacturing of the product accessories. Their causes range from label mix-up, particulate matter, and packaging issues. Though minor recalls do not threaten the health of patients directly, they are challenging to be corrected in the short-term by firms.

Table 14: Impact of market size on innovation using number of patients as proxy of market size

	(1) Trials
Log patients	3.274***
	(0.648)
residuals	−3.291***
	(0.647)
$\frac{K_{t+1}}{P_{-1}}$	1.476***
	(0.377)
average age firm	0.589***
	(0.154)
average age firm^2	−0.00478**
	(0.0016)
hhi	0.145
	(0.260)
share generics in ATC	−0.648**
	(0.244)
average age product	−0.278***
	(0.0514)
average age product^2	0.0011***
	(0.0022)
papers	0.546***
	(0.147)
papers^2	0.193
	(0.114)
# firms	0.0895***
	(0.0144)
Year Dummies	Yes
Obs.	1056
Groups	132

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered (at the ATC-3 level) standard errors are in parentheses. (1) employs MEPS database and matches the ATC-3 present in our database. Market size is measured through the number of patients within ATC-3.

Adopting the number of patients as a proxy for market size confirms the first and second stage results compared to the principal specification. The outcomes, however, turn out to be weaker in terms of significance than our main specification. Recalls do not seem a vital instrument for the number of patients. On the one hand, a more significant number of patients within an ATC-3 class might increase the probability of an adverse event than in a scarcely populated ATC-3 class, increasing the prob-

ability of a major recall. On the other hand, there is no reason to believe that an adverse event would happen in a more populated class, referring to commonly employed medicines (and therefore well tested). Moreover, a recall in a class causes a decrease in the number of patients adopting pharmaceuticals in the questioned ATC-3 class, thus compensating for the possible positive effect implied by a higher probability of adverse events. For what concerns the second stage, Tab.14 enforces the results found in Acemoglu and Linn (2004) where a coefficient of market size between 3 and 4 was found. The significance of residuals confirms the presence of endogeneity.

5 Discussion

The present Chapter provides an example of how econometrics research provides better tools than machine-learning in a complex causal setting with endogeneity and unbalanced fixed effects within a non-linear parametric model. Though machine learning provides interesting tools, it is still a novelty in many econometrics setups. Researchers must avoid its employment a priori. In the following, we sum up the main results obtained in the Chapter.

Recent research has stressed the importance of market size in determining the innovation rate in the pharmaceutical industry. At the same time, after Cerda (2007) 's critique, instrumenting with demographical shifts remains a weak, though valid, strategy. Moreover, recent contributions have stressed the importance of modeling competition and technological opportunities adequately (see Rake, 2017, Dubois et al., 2015). For example, many scholars pointed to the importance of advances in molecular biology and related fields for the industry's technological opportunities and innovative capabilities (Rake, 2017). Finally, the literature lacks aggregation analyses that easily allow drawing policy implications. For this reason, in the present Chapter, we employed ATC-3, the aggregation level used by Antitrust authorities. To provide some examples, we mention Provost et al.(2019), Markham, A. (2020), Vaishnav, A. (2011), Hawk et al.(2000), Cheng J. (2008), and other cases mostly about M&A (e.g., Case M.8889 - TEVA / PGT OTC ASSETS of 2018).

The empirical estimates are conducted on a unique database integrated with addi-

tional sources. The variety of our sources enabled us to collect and adequately classify data on trials and drug recalls in ATC-3 categories. The methodology employed is innovative (W. Lin and Wooldridge, 2019) and, differently from past techniques, permits controlling for both idiosyncratic endogeneity and heterogeneity endogeneity. The technique composes of two stages. A simple Wald test on the residuals' coefficient in the second stage allows verifying the presence of idiosyncratic endogeneity.

An innovative instrument, recalls, has been employed for the first time in literature. Recalls have been collected consulting various sources comprising FDA Enforcement reports, openFDA, and a database deriving from FOIA agreements with FDA. Recalls are representative. Major recalls have been selected to meet the criteria of sharpness, indirect effect on the dependent variable (innovation), and exogeneity. The first stage displayed a substantial (in magnitude) and significant negative impact of recalls on market size, thus validating the instrument. To the best of our knowledge, the effort constitutes an empirical novelty in the literature that mainly focuses on optimal management of recalls and provides theoretical argumentation of recalls' negative impact at the firm level. Too few papers focused on the impact of drug recalls at the market level.

Data on clinical trials have been drawn from the Clinicaltrials.gov website from the pre-clinical phase to Phase IV. They have been integrated with data on INDs from a privately owned database maintained at IMT School for Advanced Studies. To overcome issues deriving from the potential more robust response of market size to trials as a whole rather than on essential trials (i.e., bringing most probably to an innovation), only activated trials have been selected. This exercise also provides a valid answer to the argumentation that the recall of a product might imply the suspension of drug trials within its same family. Indeed, suspended and withdrawn trials have been excluded from the analysis. Nonetheless, as a robustness check, estimates are computed, including the latter in the analysis. The effort confirms the presence of idiosyncratic endogeneity and the positive sign of the estimates. However, the magnitude and the significance level decrease.

Our preferred estimates align with literature displaying an increase in the innovation of 6.3% after an increase in market size of 10%. Most recent studies of Dubois et al. (2015) display a lower coefficient of 0.23%. Authors specify how a comparison with

other works exploiting different measures of innovation remains a difficult task. They further explain that their usage of global data rather than U.S. ones for the estimations might have led to less responsiveness.

Our results are robust to several specifications. The coefficient of independent variables is in line with expectation as well as the scarce effect of lagged trials, already tested in Acemoglu and Linn (2004). Further checks confirm a positive and significant effect of market size on innovation even when fixed effects are not controlled, and the market size is considered exogenous. This latter verification partially validates (for what concerns the sign and the significance) the recent findings of Rake (2017) who did not find evidence of reverse causality. However, the coefficient's magnitude decreases sensibly compared to the preferred specification, showing a significant bias. Estimates remain robust even when no control is inserted in the analysis.

The Chapter provides exciting policy implications for what concerns innovation's stimuli and sheds some light on the impact of recalls at the market level. Governments, in particular, should be aware wherever applying either tax or price policies in the pharmaceutical sector. As already mentioned, indeed, innovation constitutes an economic phenomenon. Companies innovate mainly to have a financial return. Aware of the positive relationship between market size and innovation, authorities and policymakers should not penalize economic players too much. To guarantee citizens' future welfare, they should promote research and invest in new technologies smartly managing generics' competition.

Recalls, moreover, have not just an impact at the firm level but also the market level. Specifically, they provoke adverse market shocks, thus affecting economic stability and welfare. Authorities should therefore apply more stringent rules to avoid severe recalls. At the same time, they should consider that an intensification of Class II and Class III recalls due to the presence of more players might be physiological.

Future research might employ more up-to-date data also to include recalls of compounders and repackaging firms.

Furthermore, in the spirit of the present Dissertation mixing up econometrics and machine learning in a productive way, possible developments might employ Double Machine-Learning techniques to recover the price elasticity of demand in the pharmaceutical industry. Such information could further dig into possible relationships

among pharmaceuticals. In particular, it might be valuable to investigate the probability of recalling a drug due to the recall of another drug in the same ATC.

Chapter 4

How machine learning can boost economic intuition: application to economic complexity [1]

In this Chapter we use a machine learning technique to estimate economic complexity of countries. In contrast with the previous Chapter, the present discussion presents a traditional economic problem for which machine learning techniques are not only appropriate but also improve the theoretical debate. In the following we will introduce the problem of economic complexity and detail how machine learning has been employed to renew the literature.

Since the early 2000's, building metrics for measuring economic complexity has been a set goal. Starting from the Economic Complexity Index (ECI) developed by Hidalgo and Hausmann (2009), it has become clear how most traditional economic growth theories often shrinked internal socio-economic dynamics of countries through strict assumptions, restricting the analysis to a small subset of pre-determined factors. Differently from traditional growth theories, economic complexity measures are based

[1] This chapter is partially based on `https://assets.researchsquare.com/files/rs-1030693/v1_covered.pdf?c=1637010954` with M. Riccaboni ang G. Gnecco

on a data-driven approach, and are generally agnostic about the number and nature of factors. For instance, the ECI looks to explain the knowledge accumulated by a country and expressed in all the economic activities present in that country.
Many authors attempted to provide more and more refined measures of economic complexity. In his recent review, Hidalgo (2021) identifies two main streams of the literature on economic complexity: the first one involves metrics of so-called relatedness, whereas the second one concerns economic complexity metrics which apply dimensionality reduction techniques based, e.g., on Singular Value Decomposition (SVD). Metrics of relatedness measure the affinity between an activity and a location, while methods related to dimensionality reduction search for the best combination of factors explaining the structure of a given specialization matrix.
According to the principle of relatedness, the probability that a location c (e.g., a country) enters or exits an economic activity p (e.g., a sector) is influenced by the presence of related activities in that location. This poses, however, deeper questions about the role played by similar/nearer/other countries in determining the probability that the location c enters the economic activity p. Furthermore, while the principle of relatedness attempts to model the probability of entering an activity p by the location c, it does not provide hints about whether c will enter p successfully or not. Besides, the literature provides a strong connection between the concept of production function – a function connecting economic inputs to outputs – and economic complexity via the SVD factorization of a suitable specialization matrix $\mathbf{R}$. Indeed, SVD is used to learn the singular vectors (factors) that best explain the structure of $\mathbf{R}$. The ECI index is closely related to the leading singular vectors of that specialization matrix (Hidalgo, 2021), which provide its truncated SVD. These are also the leading eigenvectors of the product of the specialization matrix with its transpose. Normally, authors select one of the the first two eigenvectors (i.e., the ones associated with the two largest eigenvalues), because it carries out the maximum amount of information. Recently, Sciarra, Chiarotti, Ridolfi, et al. (2020b) combined information coming from the first two eigenvectors into a unique index called GENeralised Economic comPlexitY (GENEPY). Nevertheless, it should be observed that, by doing this, the other eigenvectors are neglected and together with them further information which could potentially better explain economic complexity. It looks reasonable, therefore, to ex-

plore a suitable way to select carefully also some other most informative eigenvectors beyond the first two.

The present Chapter deals with the issue above, by exploiting a class of machine-learning methods called Matrix Completion (MC). The main idea of the present work is to adopt MC to infer information about the Relative Comparative Advantage (RCA) or disadvantage of a certain country in a given class of goods or services (products). Such information is collected, for each year, in a matrix $\mathbf{RCA} \in \mathbb{R}^{C \times P}$, where C is the number of countries considered, and P is the number of products examined (at a given aggregation level). In formulas, one has

$$RCA_{c,p} := \frac{\frac{D_{c,p}}{\sum_{p'=1}^{P} D_{c,p'}}}{\frac{\sum_{c'=1}^{C} D_{c',p}}{\sum_{c'=1}^{C}\sum_{p'=1}^{P} D_{c',p'}}}, \tag{4.1}$$

where $D_{c,p}$ is the return in international dollars of the exports of country c via the product p. In case one among $D_{c,p}$, $D_{c,p'}$, $D_{c',p}$, and $D_{c',p'}$ in Eq. (4.1) is not available, one gets $RCA_{c,p} = NaN$. In this case, as a pre-processing step, that $RCA_{c,p}$ value can be replaced by 0.

Throughout the Chapter, MC is applied several times (starting from different training subsets of suitably discretized RCA values associated with several contries and products, excluding originally NaN values) to estimate the expected discretized RCA values of pairs of countries c and products p that have not been used in the training phase. To fulfill such a task, the adopted MC technique is based on a soft-thresholded SVD which selects each time – via a suitable regularization technique – the subset of most informative singular values and corresponding singular vectors. The discretization above is used to deal with the different orders of magnitude of the elements of the **RCA** matrix. The predictions provided by the several repetitions of MC are then exploited to construct two suitable surrogate incidence matrices, one of which is used to construct a novel index of economic complexity – based on MC – and the other one as input to the GENEPY algorithm Sciarra, Chiarotti, Ridolfi, et al. (2020b). In the latter case, one can investigate the differences in the output of the GENEPY when it is applied, respectively, to the original discretized **RCA** matrix and its MC surrogate.

1 Summary of the main contributions of the Chapter

The work contributes to the literature on economic complexity in three ways: (i) it applies for the first time MC to assess the complexity of countries; (ii) it defines a novel index of economic complexity based on MC; (iii) it builds up a comparison with a state-of-the-art index of economic complexity (GENEPY), revealing a high correlation between the output of GENEPY when it is applied to the original incidence matrix and the false positive rate of a binary classifier derived by the repeated application of MC. The results of our analysis show that MC performs quite well in estimating the discretized RCA values, in the sense that it allows one to discern well the case $RCA \geq 1$ from the one $0 \leq RCA < 1$. Supported by the latter evidence, we propose a novel Matrix cOmpletion iNdex of Economic complexitY (MONEY) for countries, which exploits the accuracy of their RCA predictions derived from the repeated applications of MC. Such accuracy is expressed in terms of a suitably weighted Area Under the Curve (AUC), one for each country examined. The MONEY index ranks countries according to their predictability, taking into account also the product dimension. Specifically, the larger the AUC for a specific country and the larger the average with respect to a subset of the products of that country of the MC performance in estimating the discretized RCA values of country-product pairs, the less complex that country. Using only MC to construct the proposed index helps to solve the shortcoming of GENEPY discussed above in the introduction, i.e., the fact that, differently from MC, GENEPY takes into account only the information coming from two eigenvectors. Moreover, the GENEPY index computed using the MC surrogate incidence matrix reveals interesting discrepancies in terms of economic complexity with respect to the original GENEPY, i.e., the one computed starting from the incidence matrix associated with the observed **RCA** matrix. Finally, by an analysis performed in different years (see the Supplemental), the study highlights a strong and significant positive correlation between the false positive rate of the binary classifier derived from thresholding the average output of MC (the higher the false positive rate, the more the unexpressed potential of a country in terms of activities that it could undertake) and the original GENEPY index. The false positive rate per country of

such a binary classifier is, therefore, deemed to be a suitable proxy of the original GENEPY index.

2 Proposed application of matrix completion to the RCA matrix

One of the major contributions of the present Chapter is to employ MC to study economic complexity. This class of machine-learning methods became well-known in academia after the so-called Netflix competition in the context of movie recommendation systems (see the Appendix for further details on MC and, e.g., Trevor Hastie, Mazumder, et al., 2015, Alfakih and Wolkowicz, 2000, and Cai, Candès, and Shen, 2010 for some of its applications) and was employed in this Chapter in order to estimate the expected discretized RCA values of various pairs of countries c and products p. The specific MC method adopted in the Chapter consists in completing a partially observed matrix $\mathbf{A} \in \mathbb{R}^{C \times P}$ (which is derived from the **RCA** matrix in our case), by minimizing a suitable trade-off between the reconstruction error of the known portion of that matrix and a penalty term, which penalizes a high nuclear norm of the reconstructed (or completed) matrix. This is formulated via the following optimization problem (Mazumder, Trevor Hastie, and R. Tibshirani, 2010b):

$$\underset{\mathbf{Z} \in \mathbb{R}^{C \times P}}{\text{minimize}} \left(\frac{1}{2} \sum_{(c,p) \in \Omega^{\text{tr}}} (A_{c,p} - Z_{c,p})^2 + \lambda \|\mathbf{Z}\|_* \right), \tag{4.2}$$

where Ω^{tr} is a training subset of pairs of indices (c,p) corresponding to positions of known entries of the partially observed matrix $\mathbf{A} \in \mathbb{R}^{C \times P}$, $\mathbf{Z} \in \mathbb{R}^{C \times P}$ is the completed matrix (to be optimized), $\lambda \geq 0$ is a regularization constant (chosen by a suitable validation method), and $\|\mathbf{Z}\|_*$ is the nuclear norm of the matrix $\mathbf{Z}$, i.e., the sum of all its singular values. The reader if referred to the Appendix for further technical details on the optimization problem (4.2) and on the algorithm we adopted to solve it.

While MC has already found many applications in several fields (e.g., movie recommendation, sensor engineering, economics), to the best of our knowledge, this is the first time it is employed in the context of economic complexity. More precisely, we

applied MC to define a novel complexity index. Then, we also combined it with a state-of-the-art complexity index.

In our application of MC to economic complexity, the MC optimization problem (4.2) was solved several times by a specific algorithm previously developed for that purpose (named Soft Impute Mazumder, Trevor Hastie, and R. Tibshirani, 2010b, see the Appendix), for different choices of the regularization parameter λ and of the subset Ω^{tr} (detailed later in this section). Then, two MC surrogates $\overline{\mathbf{M}}^{(MC)}$ and $\hat{\mathbf{M}}^{(MC)}$ of the incidence matrix $\mathbf{M} \in \mathbb{R}^{C \times P}$ were generated[2]. On one hand, $\overline{\mathbf{M}}^{(MC)}$ was exploited to evaluate the performance of MC by changing a suitable threshold. This allowed to build up performance measures that, combined with ad-hoc weights, compose the MONEY index (see Section 3 for details). On the other hand, $\hat{\mathbf{M}}^{(MC)}$ was used as input to the GENEPY algorithm, to construct a counterfactual $\widehat{\text{GENEPY}}^{(MC)} \in \mathbb{R}^{C \times P}$ to be compared with the GENEPY index computed using the original incidence matrix $\mathbf{M}$.

In the following, we describe our proposed way of applying MC to the reconstruction of the **RCA** matrix for the case in which the products were aggregated at the 4-digits level in the Harmonized System Codes 1992 (HS-1992). Consistently with the literature (Sciarra, Chiarotti, Laio, et al., 2018), we constructed the matrix $\mathbf{A}$ (one of the inputs to the optimization problem (4.2)) by discretizing the elements of the **RCA** matrix associated with a specific year as detailed in the Appendix. For the sake of brevity, we refer to the MC application to the definition of a measure of complexity of the countries. To get a measure of complexity of the products, it is enough to replace in the following the matrix $\mathbf{A}$ with its transpose (see also the Supplemental for some related results).

i) For the matrix $\mathbf{A} \in \mathbb{R}^{C \times P}$ (where $C = 119$ is the number of countries, and $P = 1243$ is the number of products), the MC optimization problem (4.2) was solved $N = 1000$ times by the Soft Impute algorithm, based on various choices for the training/validation/test sets (and, as already mentioned, for the regularization parameter λ).

ii) For each such repetition $n = 1, \ldots, N$, the sets above were constructed as follows.

[2]The reader is referred to the Appendix for details on how the incidence matrix $\mathbf{M}$ is defined, starting from the **RCA** matrix.

First, a (pseudo)random permutation of the rows of $\mathbf{A}$ was generated. Then, a subset S_n of these rows was considered, by including in it the first row in the permutation and the successive $s\% \simeq 25\%$ rows. In this way, the resulting number of elements of the set S_n was $|S_n| = 30$. Next, for each row in S_n, its elements belonging to all the groups except group "0" were obscured independently with probability $p_{\text{missing}} = 0.3$. The (indices ot the) remaining entries of the matrix $\mathbf{A}$ (excluding the ones belonging to the group "0") formed the training set (denoted by Ω^{tr_n}). The obscured entries in one of the $|S_n|$ rows (say, row $h \in \{1, \ldots, |S_n|\}$) formed the test set (denoted by $\Omega^{\text{test}_{n,h}}$), whereas the obscured entries in the remaining $|S_n| - 1$ rows formed the validation set (denoted by $\Omega^{\text{val}_{n,h}}$).

iii) For each repetition n, the generation of the validation and test sets from the set S_n was made $|S_n|$ times, each time with a different selection of the row h associated with the test set (and, as a consequence, also of the $|S_n| - 1$ rows associated with the validation set). Hence, the same training set was associated with $|S_n|$ different pairs of validation and test sets[3]. In this way, for each choice of S_n and of the regularization parameter λ, the MC optimization problem (4.2) was solved once instead of $|S_n|$ times, thus improving the computational efficiency. Finally, by construction, each time there was no overlap between the training, validation, and test sets.

iv) To avoid overfitting, for each choice of the training set Ω^{tr_n}, the optimization problem (4.2) was solved for 30 choices λ_k for λ, exponentially distributed as $\lambda_k = 2^{(k-1)/2}$ for $k = 1, \ldots, 30$. The resulting completed and post-processed matrix was indicated as $\mathbf{Z}^{(n)}_{\lambda_k}$. Then, for each λ_k and each of the $|S_n|$ selections of the validation sets associated with the same training set, the Root Mean Square Error (RMSE) of matrix reconstruction on that validation set was computed as

$$RMSE^{\text{val}_{n,h}}_{\lambda_k} := \sqrt{\frac{1}{|\Omega^{\text{val}_{n,h}}|} \sum_{(c,p) \in \Omega^{\text{val}_{n,h}}} \left(A_{c,p} - Z^{(n)}_{\lambda_k, c,p} \right)^2}, \tag{4.3}$$

then the choice $\lambda_{k^\circ(n,h)}$ minimizing $RMSE^{\text{val}_{n,h}}_{\lambda_k}$ for $k = 1, \ldots, 30$ was found. Finally, the RMSE of matrix reconstruction on the related test set was computed in

[3]The number of repetitions $N = 1000$ and the percentage $s\% \simeq 25\%$ were selected in order to associate each row with the test set a sufficiently large number of times, with high probability. In particular, with these choices, the average number of times each row was associated with the test set was about 250.

correspondence of the so-obtained optimal value $\lambda_{k^\circ(n,h)}$ as

$$RMSE^{\text{test}_{n,h}}_{\lambda_{k^\circ(n,h)}} := \sqrt{\frac{1}{|\Omega^{\text{test}_{n,h}}|} \sum_{(c,p)\in\Omega^{\text{test}_{n,h}}} \left(A_{c,p} - Z^{(n)}_{\lambda_{k^\circ(n,h)},c,p}\right)^2}. \tag{4.4}$$

v) For each choice of n and h, the MC predictions contained in the matrix $\mathbf{Z}^{(n)}_{\lambda_{k^\circ(n,h)}}$ were used to build a binary classifier. More precisely, each time an element $A_{c,p}$ of the matrix $\mathbf{A}$ was in the test set, such element was attributed to the class 0 (corresponding to the case $0 \leq RCA < 1$) when its MC prediction from $\mathbf{Z}^{(n)}_{\lambda_{k^\circ(n,h)}}$ was lower than 0, otherwise it was attributed to the class 1 (corresponding to the case $RCA \geq 1$). Finally, the average classification of the element $A_{c,p}$ (with respect to all the test sets to which that element belonged) was indicated as $\overline{A}^{(MC)}_{c,p} \in [0,1]$, whereas its most frequent classification (either 0 or 1) was indicated as $\hat{A}^{(MC)}_{c,p}$. A random assignment between 0 and 1 was made to deal with ties. In the (unlikely) case the element $A_{c,p}$ appeared in none of the test sets[4], both $\overline{A}^{(MC)}_{c,p}$ and $\hat{A}^{(MC)}_{c,p}$ were chosen to be equal to 0.

vi) A first MC surrogate $\overline{\mathbf{M}}^{(MC)} \in \mathbb{R}^{119\times 1243}$ of the incidence matrix $\mathbf{M}$ was defined as follows:

$$\overline{M}^{(MC)}_{c,p} \doteq \begin{cases} 0\,, & \text{if } RCA_{c,p} = NaN\,, \\ \overline{A}^{(MC)}_{c,p}\,, & \text{otherwise}\,. \end{cases} \tag{4.5}$$

Similarly, a second MC surrogate $\hat{\mathbf{M}}^{(MC)} \in \mathbb{R}^{119\times 1243}$ of the incidence matrix $\mathbf{M}$ was defined as follows:

$$\hat{M}^{(MC)}_{c,p} \doteq \begin{cases} 0\,, & \text{if } RCA_{c,p} = NaN\,, \\ \hat{A}^{(MC)}_{c,p}\,, & \text{otherwise}\,. \end{cases} \tag{4.6}$$

vii) Finally, $\overline{\mathbf{M}}^{(MC)}$ was combined with several thresholds from 0 to 1 in the first part of the construction of the proposed MONEY index (see Section 3 for details). Instead, $\hat{\mathbf{M}}^{(MC)}$ was provided as input to the GENEPY algorithm (replacing the original incidence matrix $\mathbf{M}$). In this way, a counterfactual GENEPY index, indicated as $\widehat{\text{GENEPY}}^{(MC)}$, was generated.

In order to assess the prediction capability of the binary classifier associated with

[4]Due to the choice $p_{\text{missing}} = 0.3$, each element $A_{c,p}$ not associated with the group "0" appeared in the test set on average about 75 times. So, the probability that one such element appeared in none of the test sets was negligigle.

MC (see Step 5 above), for each row (country) c of $\mathbf{A}$, we also computed the false positive rate fpr_c and the false negative rate fnr_c as the average classification error frequency, respectively, of the true negative/true positive examples in all the test sets associated with that row (where the "negative class" refers to the class 0 associated with $0 \leq RCA < 1$, and the "positive class" to the class 1 associated with $RCA \geq 1$).

3 The proposed Matrix cOmpletion iNdex of Economic complexitY (MONEY)

In this section, we introduce our proposed economic complexity index, called Matrix cOmpletion iNdex of Economic complexitY (MONEY), whose construction is based on MC.

The MONEY index is built starting from the matrix $\overline{\mathbf{M}}^{(MC)}$ introduced in Section 2. It is based on constructing a binary classifier for each country by combining the corresponding row of $\overline{\mathbf{M}}^{(MC)}$ with a threshold, then assessing the performance of the resulting MC classifications at the level of each country. First, for the binary classifier associated with each country, a Receiver Operating Characteristic (ROC) curve[5] (denoted as ROC_c) is constructed, based on a country-dependent threshold. The corresponding Area Under the Curve (AUC)[6] is denoted as AUC_c. In more details, for each country c, the elements of the c-th row of the matrix $\overline{\mathbf{M}}^{(MC)}$ are compared with a threshold to construct the associated binary classifier. The elements belonging to the same row of the original incidence matrix $\mathbf{M}$ are taken as ground

[5]We remind the reader that, for a binary classifier, the ROC curve expresses the trade-off between fall-out (false positive rate) and sensitivity (true positive rate) of that classifier, as a function of its threshold. It is recalled here that the true positive rate is equal to 1 minus the false negative rate. In general, ROC curves closer to the top-left corner indicate a better performance. As a baseline ("Bench."), a random guessing binary classifier is associated with a ROC curve with points lying along the diagonal indicated, e.g., in Fig. 8 (for which the true positive rate is equal to the false positive rate). The closer a ROC curve to the diagonal in the ROC space, the worse the performance of the associated binary classifier. It is worth reminding the reader that ROC curves do not depend on class frequencies. This makes them useful for evaluating classifiers predicting rare events as in the case of very high RCA values.

[6]We remind the reader that the AUC measures the area of the entire two-dimensional region underneath the entire ROC curve from (0,0) to (1,1). The AUC is exploited in the literature to provide an aggregate measure of performance across all possible classification thresholds. Formally, it represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, assuming that "positive" ranks higher than "negative" (Fawcett, 2006).

truth. The discrimination threshold is varied from 0 to 1, using a step size equal to 0.01. All the elements of $\overline{\mathbf{M}}^{(MC)}$ are used as dataset, except the ones having the same indices as the originally NaN values in the **RCA** matrix. This allows to form a binary classifier for each threshold and for each country. The idea now is to exploit the AUC_c of the binary classifiers associated with the countries in order to provide a measure of complexity of such countries, based on the predictability of the corresponding rows. Specifically, countries with lower AUC_c may be considered as more complex, being harder for MC to predict their RCA entries. The AUC_c alone, however, does not capture the reasons why MC performed poorly (or, vice versa, adequately). Indeed, as an example, consider the three following hypothetical scenarios. Assume that MC perfomed poorly on a country c by attributing $RCA \geq 1$ to a product p when its true RCA was smaller than 1, and assigned correctly a RCA smaller than 1 to all the other countries for the same product p (Scenario 1). Consider now the two following similar scenarios for which, for the same product p and the same country c, MC perfomed poorly on the country c by attributing $RCA \geq 1$ to the product p when its true RCA was smaller than 1, and it attributed either correctly (Scenario 2) or incorrectly (Scenario 3) $RCA \geq 1$ to all the other countries for the same product. It is reasonable to suppose that, all other things being equal, the country c to which MC assigned $RCA \geq 1$ for the product p in Scenario 1 is more complex than the same country to which MC assigned $RCA \geq 1$ for the product p in Scenarios 2 and 3. In fact, while in Scenario 2, MC could have been driven to predict, for country c, a RCA of p larger than or equal to 1 by the presence of several RCA entries larger than or equal to 1 for the other countries, this is not the case for Scenario 1. Scenario 3 is more unlikely to occur, since, as it is shown later in Section 4.1 and in the Supplemental, MC has typically a quite satisfying prediction capability in its specific application to the **RCA** matrix. In this case, it is not possible to conclude that country c is more complex than the other countries, since MC is wrongly attributing $RCA \geq 1$ to p, for all such countries.

The remarks above suggest us that, by adopting the AUC_c alone as a complexity measure, country c would be classified as equally complex in Scenarios 1, 2 and 3 (assuming the AUC_c being equal in all these cases). In order to correct for the latter bias, we propose a refined complexity measure, based on weighting the AUC_c for each

country c. The rationale of the proposed complexity measure is that not only less predictable countries (according to MC) are more complex, but one should also take into account the product dimension when comparing the MC predictions obtained for different countries, controlling for the quality of each prediction. More precisely, it is proposed to associate a weight w_c to each country c, which is constructed in such a way that the AUC_c's of countries with an higher share of "rare" false positives are weighted less (since they are less predictable). In more details, the proposed complexity measure is constructed as follows.

i) First, the MC analysis made for the countries is repeated for the products, still referring to the same year (2018). This is obtained simply by replacing at the beginning of the analysis the **RCA** matrix with its transpose. Analogously, the matrices $\hat{\mathbf{M}}^{(MC)}$ and $\overline{\mathbf{M}}^{(MC)}$ are replaced by similarly constructed matrices $(\hat{\mathbf{M}}^{\top})^{(MC)}$ and $(\overline{\mathbf{M}}^{\top})^{(MC)}$. In particular, each element of the latter matrix represents the average MC prediction for the corresponding product-country pair.

ii) Then, a threshold t is applied to the elements of the matrix $(\overline{\mathbf{M}}^{\top})^{(MC)}$. For each value of that threshold, one constructs a matrix $\left(\overline{\mathbf{M}}_t^{\top}\right)^{(MC)} \in \{0,1\}^{P\times C}$, being each entry of it equal to 1 whenever the corresponding element in the matrix $(\overline{\mathbf{M}}^{\top})^{(MC)}$ is higher than or equal to t, otherwise being it equal to 0.

iii) At this point, for each product p and each threshold t, one computes the quantity

$$ftot_{p,t} := fpr_{p,t} \times \frac{N_p}{P_p + N_p}\,, \tag{4.7}$$

being $fpr_{p,t}$ the false positive ratio for the classifications associated with that product (determined by the comparison between $\left(\overline{\mathbf{M}}_t^{\top}\right)^{(MC)}$ and $\mathbf{M}^{\top}$, restricted to the entries associated with that product) and $\frac{N_p}{N_p+P_p}$ the proportion of entries with true $RCA < 1$ with respect to all the entries associated with that product (i.e., 119). Besides, the average $\overline{ftot}_p$ of $ftot_{p,t}$ with respect to t is computed.

iv) Then, for each country c, the weight w_c is defined as follows:

$$w_c := \frac{\sum_{p=1}^{P}(\hat{M}^{\top})_{p,c}^{(MC)} \times \overline{ftot}_p}{\sum_{p=1}^{P}(\hat{M}^{\top})_{p,c}^{(MC)}}\,. \tag{4.8}$$

In other words, for each country c, the weight w_c is the average of $\overline{ftot}_p$ with respect to all the products p for which one predicts $RCA \geq 1$ through the surrogate incidence matrix $(\hat{\mathbf{M}}^{\top})^{(MC)}$.

v) Finally, the MONEY index for each country c is computed as:

$$MONEY_c := 1 - w_c \times AUC_c \,. \tag{4.9}$$

4 Results

4.1 Global performance of matrix completion

In the following, the diagnostic ability of MC is illustrated. Likewise in Section 3, the matrix $\overline{\mathbf{M}}^{(MC)}$ was combined with a threshold to construct a binary classifier (in this case, however, differently from Section 3, the threshold did not depend on the country). The discrimination threshold was varied from 0 to 1, using a step size equal to 0.01. All the elements of $\overline{\mathbf{M}}^{(MC)}$ were used as dataset, except the ones having the same indices as the originally NaN values in the **RCA** matrix. The ground truth was provided by the corresponding elements of the original incidence matrix $\mathbf{M}$. Fig. 8 shows the resulting ROC curve. Similarly, ROC_c curves (see Section MONEY) for a random sample of countries are displayed in Fig. 9.

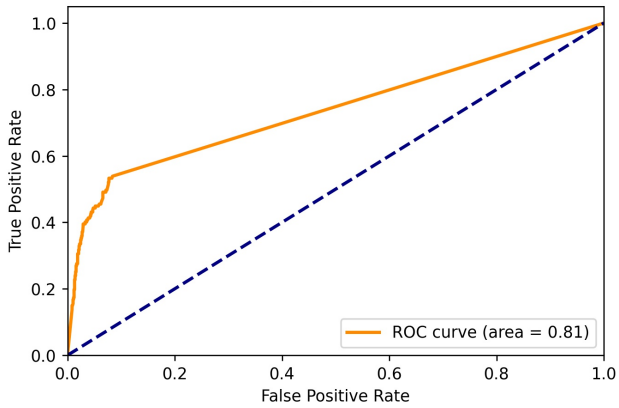


Figure 8: Global ROC curve constructed starting from the matrix $\overline{\mathbf{M}}^{(MC)}$, for the year 2018. "Bench." stands for the line passing through the origin with slope 1.

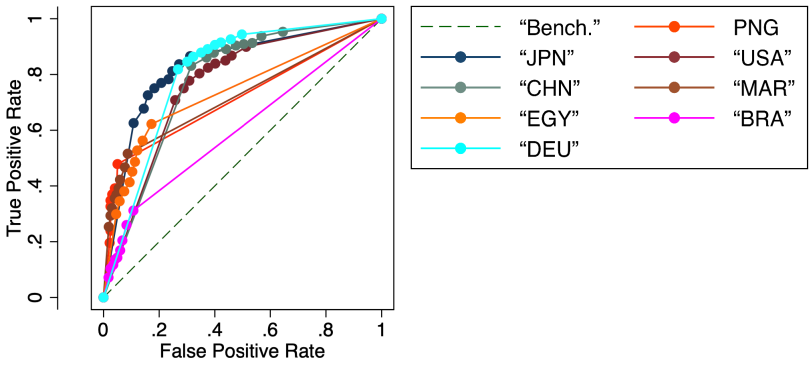
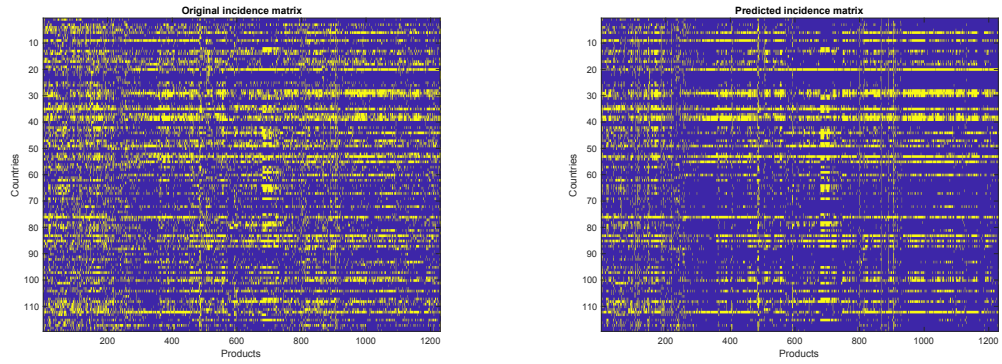


Figure 9: ROC_c curves constructed starting from the matrix $\overline{\mathbf{M}}^{(MC)}$, for a sample of countries and for the year 2018. "Bench." stands for the line passing through the origin with slope 1.

As it is evident from Figs. 8 and 9, MC performed quite well on average both globally and for developed countries such as Japan, United States and Germany. Its performance was poorer (though still above the baseline) for countries that either provided less information on their trade flows or whose trade flows were extremely volatile (i.e., they alternated between products with extremely high RCA values and products with very low RCA values). Specifically, fnr_c was higher for the latter countries. Nonetheless, the average performance of MC over all the countries was high as depicted by the AUC reported in Fig. 8, which turned out to be about 0.81 for the binary classifier described in Step 5 of Section 2.

As a further check, since the positive and negative labels were unbalanced in the original dataset (specifically, entries with $RCA < 1$ represented almost the 70% of the entire dataset), we also applied the Balanced Accuracy (BACC) index[7], which turned out to be 0.75. Figs. 10(a)-10(b) display the original incidence matrix $\mathbf{M}$ as compared to the MC surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$ obtained at the HS-4 level of product aggregation. The two matrices display similar but not identical entries. On one hand, their similarity confirms the good MC prediction performance at a global level. On the other hand, their differences could be attributed to the high complexity of specific country/product pairs being predicted. In other words, there may be a discrepancy between the actual RCA value of a country/product pair and

[7]The BACC is a performance metric designed for binary classifiers in the case of unbalanced datasets. It is calculated as the average of the proportion of correctly classified elements of each class individually, and ranges from 0 (low balanced accuracy) to 1 (maximum balanced accuracy). Formally, it is equal to $(tpr + tnr)/2$, where t stands for "true".



Original incidence matrix $\mathbf{M}$. Surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$.

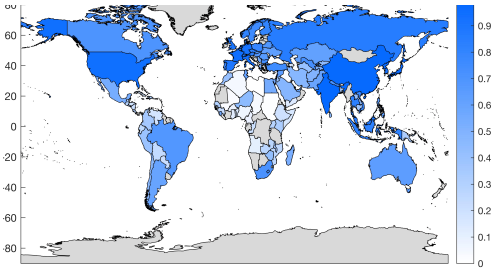
Figure 10: Original versus surrogate incidence matrix for the year 2018 at the HS-4 level.

its potential RCA value, predicted by MC on the basis of similar country/product pairs.

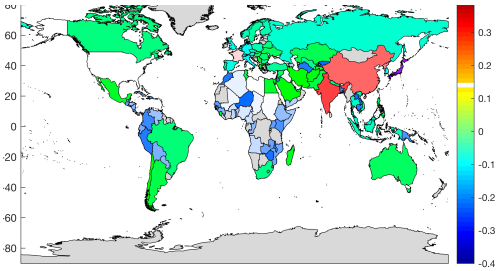
4.2 Results related to the MONEY index

In this section we report the ranking of countries in terms of economic complexity as expressed by the MONEY index introduced in Section 3. In particular, we represent the countries according to their MONEY index (Fig.11(a)), then we compare the obtained ranking with the one expressed by GENEPY (Fig.11(b)). In Fig. 11(a), countries are colored according to their MONEY values (normalized between 0 and 1), which are proportional to the shade of blue. In particular, the color map ranges from the least complex countries c (colored in white) to the most complex ones (colored in dark blue).

It is worth observing that both the GENEPY and the proposed MONEY index arise from the attempt to reconstruct (in a different way for each method) a matrix re-



Countries colored according to the MONEY index. Countries in darkest shades of blue are associated with a lower MONEY, hence they are considered more complex. Countries colored in grey are not considered in the analysis.



Countries colored according to the difference between the values of their MONEY and GENEPY indices. Countries in darker shades of blue are associated with a value of MONEY<GENEPY, vice versa countries in red.

Figure 11: Values of the MONEY index for the year 2018 at the HS-4 level of aggregation and their differences with respect to the corresponding values of the GENEPY index for the same year and the same level of aggregation.

lated to trade flows. In the case of GENEPY, the matrix is a proximity matrix **N** derived from the incidence matrix **M** (see the Appendix for the definition of the matrix **N**), and its reconstruction is obtained as a nonlinear least-square estimate based on the components of the first two (normalized) eigenvectors of that matrix. Then, a successive evaluation on how the quality of the estimate changes by dropping specific components of such eigenvectors (the ones associated with a given country) is made. In our case, the matrix **A** is obtained as a discretization of the **RCA** matrix. Then, MC is applied several times to the matrix **A** to reconstruct a portion of that matrix which has been obscured, in the attempt to uncover a "latent" similarity between countries, which can be useful for the prediction of whether their RCA entries are lower than 1, or higher than or equal to 1. Another difference is that the matrix reconstruction on which GENEPY is based relies only on two eigenvectors of **N**, whereas our method, being also based on MC, exploits a typically much larger number of left-singular/right-singular vectors to build the reconstructed matrix, for each application of MC. The choice of the number of such pairs is made automatically by the adopted validation procedure. Moreover, a final evaluation of the quality of the reconstruction is made, by considering several test sets, on which the AUC_c's are based. A further quality assessment is provided by Tab. 15, which reports the

number of G19+5[8] countries in the top 20, 30 and 40 positions, computed according to each among the GENEPY, $\widehat{\text{GENEPY}}^{(MC)}$ and MONEY indices, then divided by 24. It is evident from the table that the largest ratio is obtained in correspondence of the proposed MONEY index. Additional robustness checks related to the MONEY index computed in years different than 2018 are reported in the Supplemental.

Table 15: Number of G19+5 countries in the top 20, 30 and 40 positions for the year 2018, computed according to each among the GENEPY, $\widehat{\text{GENEPY}}^{(MC)}$ and MONEY indices, then divided by 24.

	# of G19+5 countries in the top x positions/24		
Index	$x = 20$	$x = 30$	$x = 40$
GENEPY	42%	58%	75%
$\widehat{\text{GENEPY}}^{(MC)}$	55%	64%	71%
MONEY	55%	66%	79%

MONEY and GENEPY indices are also produced at the HS-2 level in the Supplemental part. To provide a further contribution to the literature on how the granularity of the products changes the knowledge on countries' complexity, below we include a broad investigation on how the MONEY and GENEPY rankings differ between the two aggregations.

The correlation between MONEY and GENEPY at the HS-4 level is listed below:

$$Kendall's \quad \tau = 0.802;$$

$$Spearrman's \quad \rho = 0.818.$$

The correlation between MONEY and GENEPY at the HS-2 level is shown below:

$$Kendall's \quad \tau = 0.831;$$

$$Spearrman's \quad \rho = 0.845.$$

As detailed above, the correlation between MONEY and GENEPY is higher at HS-2 level than at HS-4.

We guess that, on one side, at HS-2, matrix completion is facilitated in reconstructing

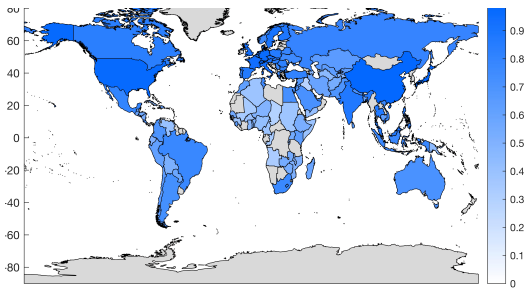
[8]In the table we considered countries within G20. However, since G20 countries comprise EU (except France, Italy and Germany, which are accounted separately), that is an agglomerate of countries, we considered a group of 5 representative countries for EU, namely: Spain, Switzerland, Greece, Denmark and Hungary.

the original matrix. On the other side, the eigenvalue problem of GENEPY is much simplified (concerning considerably fewer equations than the original HS-4 matrix). Since both the approaches face a simplified problem, and since the number of eigenvalues taken by Matrix Completion is not reduced proportionally in this exercise, it might be that the additional information provided by the fewer eigenvalues selected through Matrix Completion in the HS-2 case is better summarised by the first two eigenvalues (taken by GENEPY) than in the HS-4 case.

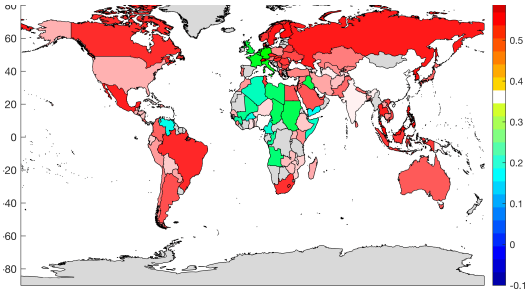
4.3 Differences in the GENEPY indices based on the original incidence matrix M and on $\hat{\mathbf{M}}^{(MC)}$

Fig. 12(a) represents the GENEPY index computed based on the original incidence matrix **M**. The interpretation is the same as in Fig.11(a). It is worth noticing that the GENEPY value computed based on the incidence matrix **M** and the $\widehat{\text{GENEPY}}^{(MC)}$ value based on its surrogate $\hat{\mathbf{M}}^{(MC)}$ are quite similar (see Fig. 12(b) for their difference). Hence, they provide analogous results in terms of the complexity of the countries, confirming the satisfactory prediction capability of MC for the specific learning task. Nevertheless, one can also notice that the two complexities differ in some countries. Such differences may be ascribed to surpluses/deficits of the actual complexities of such countries (i.e., the ones measured by GENEPY based on the original incidence matrix **M**) with respect to the respective predicted complexities (i.e., the ones measured by $\widehat{\text{GENEPY}}^{(MC)}$, which is based on the surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$).

To quantify the correlation between the GENEPY rankings computed based on **M** and $\hat{\mathbf{M}}^{(MC)}$, respectively, we evaluated their Kendall rank correlation coefficient τ_k. The statistical test produced $\tau_k \simeq 0.8$ with a p-value near 0, rejecting significantly the null hypothesis of independence between GENEPY and $\widehat{\text{GENEPY}}^{(MC)}$.



Countries colored according to the GENEPY index computed starting from the original incidence matrix $\mathbf{M}$ at the HS-4 level of aggregation.



Difference between the GENEPY index in (a) and the $\widehat{\text{GENEPY}}^{(MC)}$ index based on its surrogate $\hat{\mathbf{M}}^{(MC)}$.

Figure 12: Original GENEPY values for countries for the year 2018 at the HS-4 level of aggregation, and their comparison with their $\widehat{\text{GENEPY}}^{(MC)}$ values. Countries colored in grey are not considered in the analysis.

It is worth noticing that, with a few exceptions (China, France, Italy, UK and Germany) the more complex the country according to GENEPY, the higher the difference between GENEPY and $\widehat{\text{GENEPY}}^{(MC)}$. Finally, Fig. 13 displays the false positive rate fpr_c for each country considered in the analysis, which turned out to produce a ranking of countries quite similar to the one generated by GENEPY ($\tau_k = 0.75$). Notice that the correlation between false positive and GENEPY is 0.9636 at HS-4 level and 0.9496 at the HS-2 level.

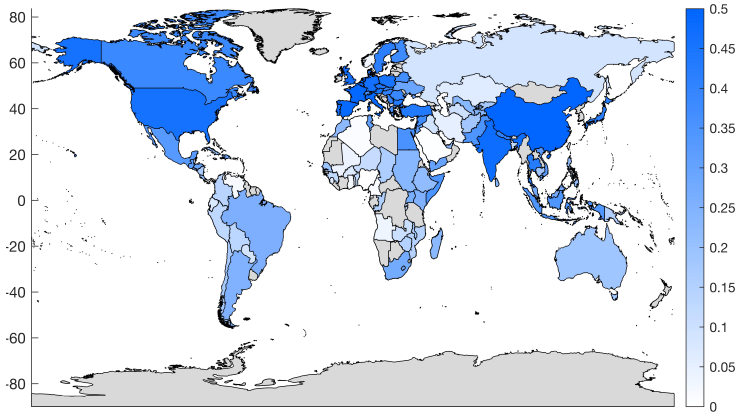


Figure 13: False positive rate fpr_c, reported proportionally to the shade of blue for the year 2018 and the product aggregation level HS-4.

5 Discussion

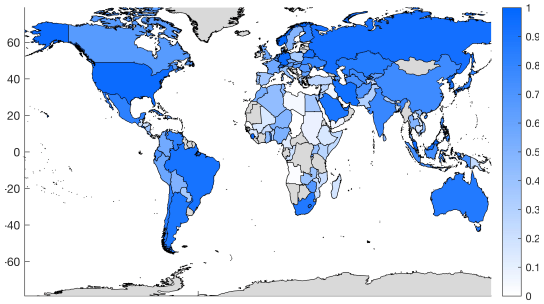
In the present Chapter, we applied Matrix Completion (MC) to investigate in various ways the economic complexity of countries. First, we assessed a quite high accuracy of the MC predictions, when MC was applied to reconstruct the Revealed Comparative Advantage (RCA) matrix, which is at the basis of the construction of several existing economic complexity indices (see the Appendix). Then, we proposed the Matrix cOmpletion iNdex of Economic complexitY (MONEY), based on the predictability of the RCA entries associated with different countries. As an additional contribution, we combined MC with a recently-developed economic complexity index (GENEPY), to assess the expected economic complexity of countries. In the Chapter, MC was exploited to infer the expected discretized RCA of a country c in a certain class of goods or services p. The MC technique employed is based on a soft-thresholded SVD. This, combined with the MC validation phase, allows to select automatically a suitable number of singular vectors to be used to reconstruct the discretized **RCA** matrix. In this way, differently from previous economic complexity indices, the information extracted is not restricted to the first two singular vectors.

The results of our analysis highlighted a generally quite good performance of MC in discerning country-product pairs with RCA values greater than or equal to the critical threshold of 1, denoting the competitiveness of c in producing p. The outcomes were summarized by reporting the global ROC curve and comparing the heat-map of the true incidence matrix $\mathbf{M}$ and the one of its MC surrogate matrix $\hat{\mathbf{M}}^{(MC)}$, which was obtained from various applications of MC. Motivated by the high MC accuracy, we developed the MONEY index taking into account both the predictive performance of MC for each country (as measured by its AUC_c) and the product dimension. In other words, when constructing that index, each AUC_c was weighted by the average of the $\overline{ftot}_p$'s with respect to a subset of products associated with the specific country. As a further step, we applied the GENEPY algorithm first to the incidence matrix derived directly from the original **RCA** matrix, then to the MC surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$. This allowed us to directly compare the values of the two GENEPY indices, thus assessing their potential discrepancies. On average, such discrepancies were higher for more complex countries according to the original GENEPY index.

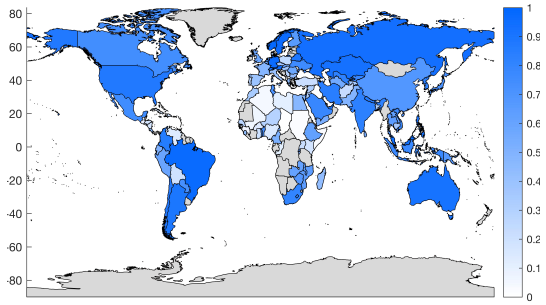
Supplemental

MONEY index for the years 2005 and 2014

The present section provides robustness checks on the MONEY index. In particular, the whole analysis has been repeated for the years 2005 and 2014. For such years, the global AUC (see Section 4.1) is respectively 0.76 for 2005 and 0.80 for 2014. For the same reason as reported in the main text – that is, the fact that the two categories $RCA > 1$ and $0 \leq RCA < 1$ are unbalanced – we also computed the BACC. The latter amounts to 0.74 for 2005 and to 0.75 for 2014.
Figures 14(a)-14(b) display the MONEY index for the years 2005 and 2014.



Values of the MONEY index for the year 2005 at the HS-4 level of aggregation.



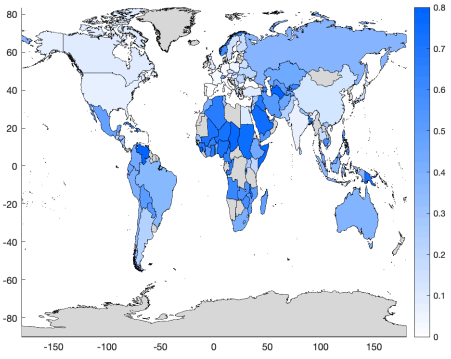
Values of the MONEY index for the year 2014 at the HS-4 level of aggregation.

Figure 14: Original MONEY values for countries for the year 2005 and 2014 at the HS-4 level of aggregation. Countries colored in grey are not considered in the analysis.

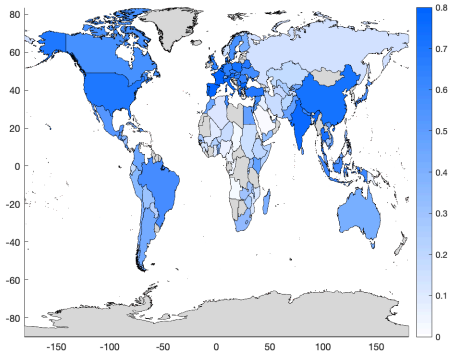
The figures display similar results to those obtained in the main analysis. However, some differences emerge. Specifically, Russia and East Asia appear to be more complex in 2005 and 2014 than in 2018. Such countries were indeed rapidly developing in those years and their growth rate was higher than the one of EU, which was slowed down by the 2008 financial crisis. The crisis had an impact also on USA. Its MONEY index was in fact lower in 2014 as compared to 2005. In the successive years, yet, EU and USA recovered from the crisis, and their complexity, as measured by the MONEY index, raised accordingly in 2018.

Application of the second part of the analysis to countries for the year 2018, with products aggregated at the HS-2 level

Fig. 15 reports, for the product aggregation level HS-2, results similar to those obtained in Section 4.3 of the main text for the HS-4 level. For the sake of completeness, we report also results for the false negative rate. Fig. 16 provides similar results, restricting to the countries for which both the false negative rate and the false positive are lower than 0.5.



False negative rate fnr_c, reported proportionally to the shade of blue. Countries colored in grey are not considered in the analysis.



False positive rate fpr_c, reported proportionally to the shade of blue. Countries colored in grey are not considered in the analysis.

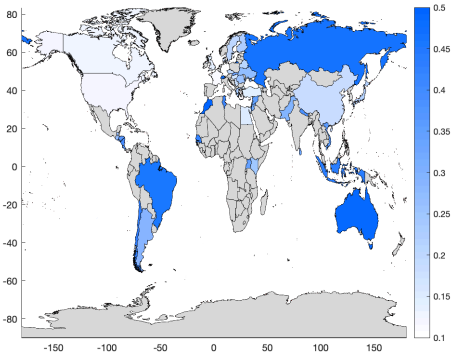
Figure 15: False negative and false positive rates for countries, obtained by the method reported in Step 5 of Section 2 for the year 2018 and the HS-2 level of aggregation.

Additionally, Tab. 16 reports, for the product aggregation level HS-2, the Kendall correlation coefficients τ_k between the ranking produced using GENEPY against the ones produced using either $fnr_{c,hs-2}$ or $fpr_{c,hs-2}$.

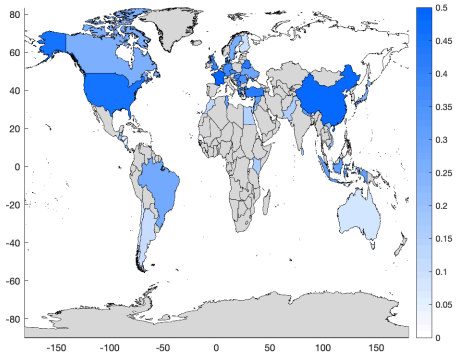
	GENEPY (τ_k)	**GENEPY** (p-**value**)
$fnr_{c,hs-2}$	0.1230	0.0575
$fpr_{c,hs-2}$	0.6476	0.0000

Table 16: Kendall rank correlation coefficients τ_k and corresponding p-values for the 2018 ranking of countries based on the HS-2 level of aggregation and produced using GENEPY against the 2018 rankings produced respectively by fnr_c, and fpr_c.

Similarly, Figs. 17(a)-17(b), which refer to products aggregated at the HS-2 level,



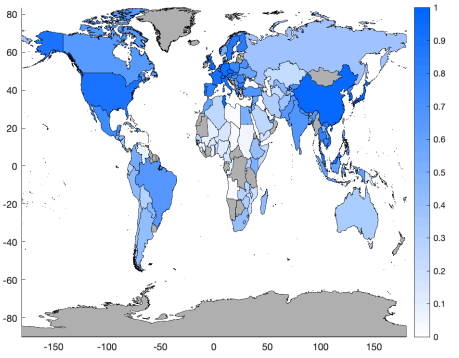
False negative rate fnr_c for countries c having both average false negative and false positive rates lower than 0.5. In the figure, fnr_c is reported proportionally to the shade of blue. Countries colored in grey are not considered in the analysis.



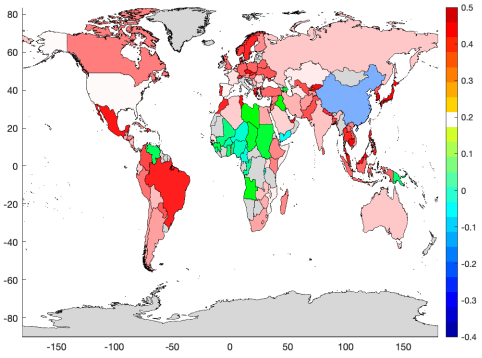
False positive rate fpr_c for countries c having both average false negative and false positive rates lower than 0.5. In the figure, fpr_c is reported proportionally to the shade of blue. Countries colored in grey are not considered in the analysis.

Figure 16: False negative and false positive rates for a selection of countries, obtained by the method reported in Step 5 of Section 2 for the year 2018 and the HS-2 level of aggregation.

show results similar to those obtained at product level HS-4.



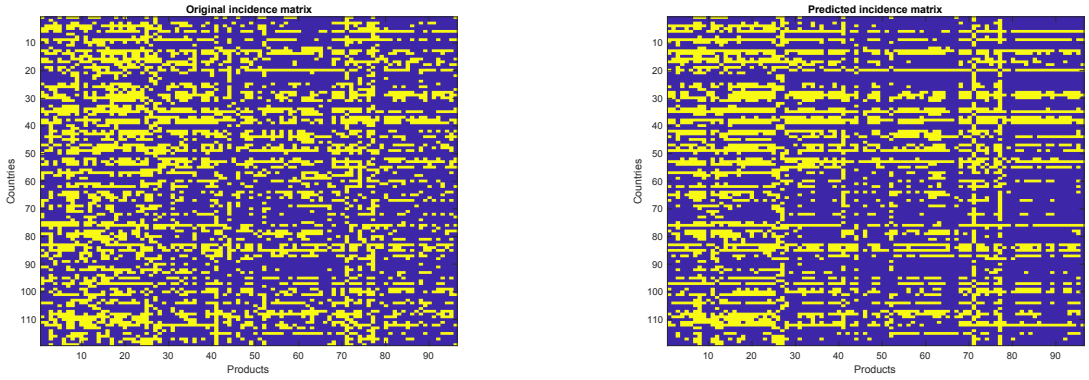
Countries colored according to the GENEPY index computed starting from the original incidence matrix $\mathbf{M}$. Countries colored in grey are not considered in the analysis.



Difference between the GENEPY indices computed starting from the original incidence matrix $\mathbf{M}$ and on its surrogate $\hat{\mathbf{M}}^{(MC)}$. Countries colored in grey are not considered in the analysis.

Figure 17: Original GENEPY values for countries for the year 2018 with products aggregated at the HS-2 level, and their comparison with their $\widehat{\text{GENEPY}}^{(MC)}$ values.

Additionally, Figs. 18(a)-18(b) report the original incidence matrix $\mathbf{M}$ as compared to its MC surrogate $\hat{\mathbf{M}}^{(MC)}$ obtained at the HS-2 level of product aggregation. Also in this case, the two matrices display similar but not identical entries. Thus, similar conclusions to the ones obtained for the HS-4 case apply. However, for the HS-2 case, the percentage of elements in which the two matrices differ is lower.



Original incidence matrix $\mathbf{M}$.

Surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$.

Figure 18: Original versus surrogate incidence matrix for the year 2018 and the HS-2 level of product aggregation.

We report below, some error metrics between $\mathbf{M}$ and $\hat{\mathbf{M}}^{(\mathbf{MC})}$ in the HS-2 case (see Fig. 18(a) and 18(b)) and in the HS-4 case (Fig. 10(a) and 10(b)). Specifically, the error probability taken as the sum of logical symmetric differences of elements of $\mathbf{M}$ and $\hat{\mathbf{M}}^{(\mathbf{MC})}$ for both HS-2 and HS-4 levels divided by the matrix size (both $\mathbf{M}$ and $\hat{\mathbf{M}}^{(\mathbf{MC})}$ have the same number of elements). In particular,

$$P_{err,HS-2} = 0.21;$$
$$P_{err,HS-4} = 0.14;$$

Spearman correlation

$$\rho_{HS-2} = 0.52;$$
$$\rho_{HS-4} = 0.53.$$

To conclude, Tab. 17 reports the detail of the differences $\text{GENEPY} - \widehat{\text{GENEPY}}^{(MC)}$ for countries, for the year 2018 at the HS-2 level of aggregation. Such differences were already analysed in Fig.17(b).

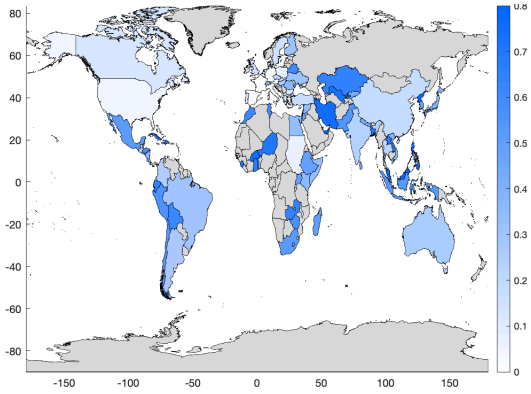
GENEPY-$\widehat{\text{GENEPY}}^{(MC)}$	Country
-0.39257	Hong Kong
-0.32742	China
-0.11648	Italy
0.012212	Iraq
0.027759	Libyan Arab Jamahiriya
0.029027	Angola
0.045068	Azerbaijan
0.047551	Chad
0.054625	Sudan
0.061237	Guinea
0.063626	Congo
0.064162	Venezuela
0.07725	Ghana
0.080608	Papua New Guinea
0.081877	Mali
0.083468	Benin
0.08376	Congo
0.084289	Sierra Leone
0.089644	Burkina Faso
0.089985	Nigeria
0.090831	Burundi
0.092596	Niger
0.094417	Cameroon
0.096485	Rwanda
0.096501	Yemen
0.097521	United States
0.09755	South Sudan
0.099289	Spain
0.10046	Mozambique
0.10293	Somalia
0.10471	Turkmenistan
0.10538	Kazakhstan
0.10658	Senegal
0.10809	Ivory Coast
0.10911	France
0.11116	Bulgaria
0.11269	Malawi
0.11666	Uganda
0.11966	Romania
0.12069	Saudi Arabia
0.12078	Netherlands
0.12121	Algeria
0.12298	Ecuador
0.12453	Russia
0.12549	Australia
0.12928	Zambia
0.13043	India
0.13344	Nicaragua
0.13778	Bolivia
0.14032	Serbia
0.14037	Togo
0.14076	Iran Islamic Republic of
0.14421	Zimbabwe
0.1452	Belgium
0.14975	Guatemala
0.15105	Egypt
0.1534	South Africa
0.15421	Chile
0.15709	Cuba
0.16116	Argentina

GENEPY-$\widehat{\text{GENEPY}}^{(MC)}$	Country
0.16381	Afghanistan
0.16641	Haiti
0.16862	Tajikistan
0.16915	Lebanon
0.17253	Austria
0.17402	Finland
0.17561	Honduras
0.17755	Madagascar
0.17791	Ethiopia
0.18072	Viet Nam
0.18157	Kenya
0.19125	Syrian Arab Republic
0.19381	Canada
0.19577	Jordan
0.19605	Costa Rica
0.19724	Peru
0.19725	Lao People's Democratic Republic
0.19882	United Kingdom
0.19962	Indonesia
0.201	Turkey
0.20222	Hungary
0.20318	Poland
0.20356	Germany
0.20357	Belarus
0.2047	Korea, Democratic People's Republic of
0.20659	Sri Lanka
0.20667	Paraguay
0.20859	Ukraine
0.21088	Pakistan
0.21494	Uzbekistan
0.21626	Israel
0.21999	Denmark
0.2241	Colombia
0.22585	Tanganyika
0.22832	Bangladesh
0.2353	Singapore
0.23852	Norway
0.23875	Switzerland
0.23978	Dominican Republic
0.242	Republic of the Union of Myanmar
0.24262	El Salvador
0.24271	Nepal
0.25116	Thailand
0.2632	Portugal
0.26521	Sweden
0.26842	Morocco
0.27857	Brazil
0.28289	Philippines
0.28441	Slovakia
0.30354	United Arab Emirates
0.30412	Greece
0.3389	Mexico
0.35332	Tunisia
0.35893	Kyrgyzstan
0.37339	Malaysia
0.38045	Czech Republic
0.39382	Cambodia
0.50948	Japan
0.5384	Korea, Republic of

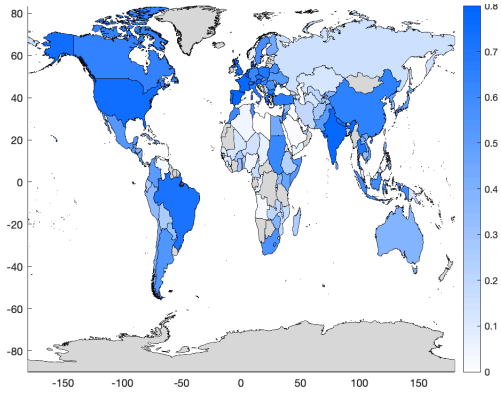
Table 17: Countries classified according to the deviations GENEPY$-\widehat{\text{GENEPY}}^{(MC)}$ for the year 2018 at the HS-2 level. Countries colored in blue are countries for which GENEPY$<\widehat{\text{GENEPY}}^{(MC)}$. Countries colored in green are countries for which GENEPY$\simeq\widehat{\text{GENEPY}}^{(MC)}$. Countries colored in red are countries for which GENEPY$>\widehat{\text{GENEPY}}^{(MC)}$.

Results of the analysis for countries, for the years 2005 and 2014 and at the HS-2 level

In the following, results similar to those obtained in the main text are reported for the years 2005 and 2014. In this case, in order to reduce the computational effort, the analysis was made at the HS-2 level of product aggregation. Figs. 19-20 report false negative and false positive rates for the two years, whereas Tab.18 reports Kendall correlation coefficients between the ranking produced using GENEPY against the rankings produced respectively by $fnr_{c,t}$ and $fpr_{c,t}$, for the years $t = 2005$ and $t = 2014$.

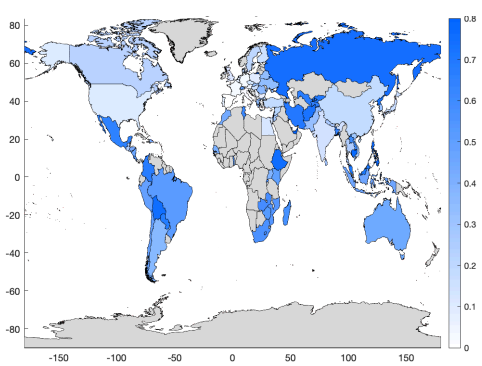


False negative rate fnr_c, reported proportionally to the shade of blue for the year 2005. Countries colored in grey are not considered in the analysis.

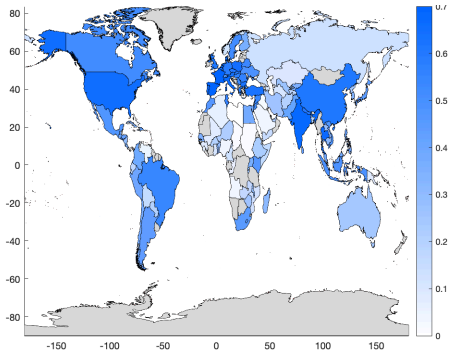


False positive rate fpr_c, reported proportionally to the shade of blue for the year 2005. Countries colored in grey are not considered in the analysis.

Figure 19: False negative and false positive rates for countries, obtained by the method reported in Step 5 of Section 2 for the year 2005, with products aggregated at the HS-2 level.



False negative rate fnr_c, reported proportionally to the shade of blue for the year 2014. Countries colored in grey are not considered in the analysis.



False positive rate fpr_c, reported proportionally to the shade of blue for the year 2014. Countries colored in grey are not considered in the analysis.

Figure 20: False negative and false positive rates for countries, obtained by the method reported in Step 5 of Section 2 for the year 2014, with products aggregated at the HS-2 level.

	GENEPY$_t$ (τ_k)	**GENEPY**$_t$ (p-**value**)
$fnr_{c,2005}$	0.2018	0.0020
$fpr_{c,2005}$	0.6212	0.0000
$fnr_{c,2014}$	0.1936	0.0031
$fpr_{c,2014}$	0.6165	0.0000

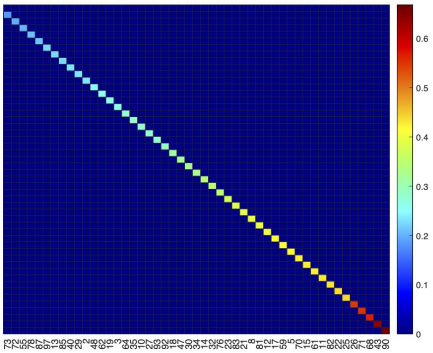
Table 18: τ_k and relative p-values for the rankings produced using GENEPY against the ranking produced respectively by $fnr_{c,t}$ and $fpr_{c,t}$, for the years $t = 2005$ and $t = 2014$, with products aggregated at the HS-2 level.

Application of the analysis to the products at the HS-2 level

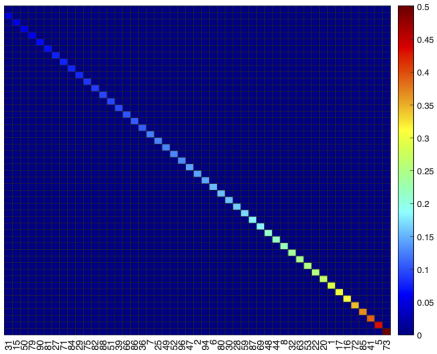
The same analysis made in the main text for the countries has been repeated for the products, still referring to the year 2018. This is obtained simply by replacing at the beginning of the analysis the **RCA** matrix with its transpose. Notice that this analysis, as some of the analyses reported in this Supplemental, was made at the HS-2 level for computational time reasons. The results obtained at the HS-2 level, however, correlated at the 95% with the ones obtained at the HS-4 level.
Figs. 21 displays respectively, on the main diagonal of each matrix reported, and for a subset of product codes,

- the false negative rate fnr_p for each product p (Fig.21(a));
- the false positive rate fpr_p for each product p (Fig.21(b));
- the false negative rate fnr_p, for the subset of products p for which both fnr_p and fpr_p are lower than 0.5 (Fig.21(c));
- the false positive rate fpr_p, for the subset of products p for which both fnr_p abd fpr_p are lower than 0.5 (Fig.21(d)).

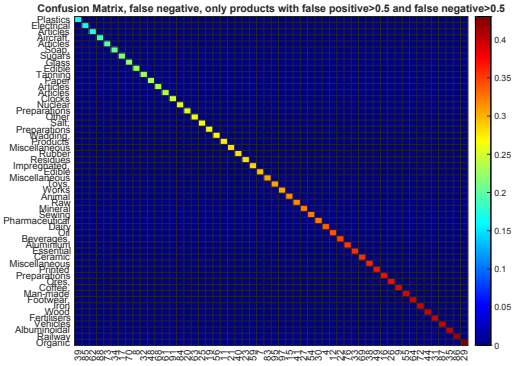
In all these cases, the products have been ordered increasingly with respect to the (either false positive or false negative) rate.



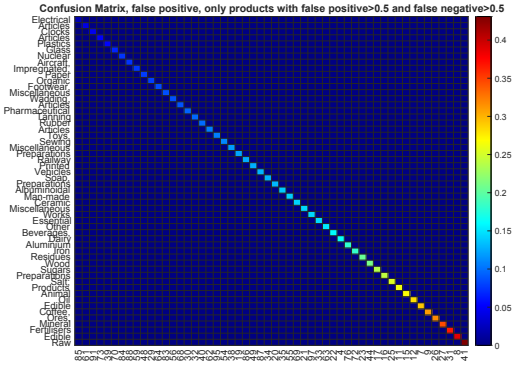
False negative rate fnr_p for the products, reported as a function of the color shade from blue (associated with the lowest value) to red (associated with the highest value).



False positive rate fpr_p for the products, reported as a function of the color shade from blue (associated with the lowest value) to red (associated with the highest value)



False negative rate fnr_p for the products having both $fnr_c < 0.5$ and $fpr_c < 0.5$. Values of fnr_c are reported as a function of the color shade from blue (associated with the lowest value) to red (associated with the highest value).



False positive rate fpr_p for the products having both $fnr_c < 0.5$ and $fpr_c < 0.5$. Values of fpr_c are reported as a function of the color shade from blue (associated with the lowest value) to red (associated with the highest value).

Figure 21: False negative and false positive rates for products, obtained by the method reported in Step 5 of Section 2 for the year 2018 and the HS-2 level.

The correspondence between product codes and their names is reported in Tab. 19. The names are reported only for the HS-2 level for a better readability.

Row index	Product name
1	Live animals
2	Meat and edible meat offal
3	Fish and crustaceans, molluscs and other aquatic invertebrates
4	Dairy produce
5	Products of animal origin, not elsewhere specified or included
6	Live trees and other plants
7	Edible vegetables and certain roots and tubers
8	Edible fruit and nuts;
9	Coffee, tea, mate and spices
10	Cereals
11	Products of the milling industry;
12	Oil seeds and oleaginous fruits;
13	Lac; gums, resins and other vegetable saps and extracts
14	Vegetable plaiting materials
15	Animal or vegetable fats and oils and their cleavage products
16	Preparations of meat, of fish or of crustaceans
17	Sugars and sugar confectionery
18	Cocoa and cocoa preparations
19	Preparations of cereals, flour, starch or milk; pastrycooks' products
20	Preparations of vegetables, fruit, nuts or other parts of plants
21	Miscellaneous edible preparations
22	Beverages, spirits and vinegar
23	Residues and waste from the food industries; prepared animal fodder
24	Tobacco and manufactured tobacco substitutes
25	Salt; sulphur
26	Ores, slag and ash
27	Mineral fuels, mineral oils and products of their distillation; bituminous substances
28	Inorganic chemicals; organic or inorganic compounds of precious metals, of rare-earth metals, of radioactive elements or of isotopes
29	Organic chemicals
30	Pharmaceutical products
31	Fertilisers
32	Tanning or dyeing extracts
33	Essential oils and resinoids
34	Soap, organic surface-active agents, washing preparations, lubricating preparations, artificial waxes...
35	Albuminoidal substances
36	Explosives, fireworks
37	Photographic or cinematographic goods
38	Miscellaneous chemical products
39	Plastics and articles thereof
40	Rubber and articles thereof
41	Raw hides and skins
42	Articles of leather; saddlery and harness
43	Furskins and artificial fur
44	Wood and articles of wood
45	Cork and articles of cork
46	Manufactures of straw, of esparto or of other plaiting materials
47	Pulp of wood or of other fibrous cellulosic material
48	Paper and paperboard

Row index	Product name
49	Printed books, newspapers, pictures and other products of the printing industry
50	Silk
51	Wool, fine or coarse animal hair
52	Cotton
53	Other vegetable textile fibres
54	Sewing thread of man-made filaments
55	Man-made staple fibres
56	Wadding, felt and nonwovens
57	Carpets and other textile floor coverings
58	Special woven fabrics
59	Impregnated, coated, covered or laminated textile fabrics
60	Knitted or crocheted fabrics
61	Articles of apparel and clothing accessories, knitted or crocheted
62	Articles of apparel and clothing accessories, not knitted or crocheted
63	Other made up textile articles
64	Footwear, gaiters and the like; parts of such articles
65	Headgear and parts thereof
66	Umbrellas, sun umbrellas and similar articles
67	Prepared feathers and down and articles made of feathers or of down
68	Articles of stone, plaster, cement, asbestos, mica or similar materials
69	Ceramic products
70	Glass and glassware
71	Natural or cultured pearls, precious or semi-precious stones, precious metals
72	Iron and steel
73	Articles of iron or steel
74	Copper and articles thereof
75	Nickel and articles thereof
76	Aluminium and articles thereof
78	Lead and articles thereof
79	Zinc and articles thereof
80	Tin and articles thereof
81	Other base metals; cermets; articles thereof
82	Tools, implements, cutlery, spoons and forks, of base metal; parts thereof of base metal
83	Miscellaneous articles of base metal
84	Nuclear reactors, boilers, machinery and mechanical appliances
85	Electrical machinery and equipment and parts thereof
86	Railway or tramway locomotives, rolling-stock and parts thereof
87	Vehicles other than railway or tramway rolling-stock, and parts and accessories thereof
88	Aircraft, spacecraft, and parts thereof
89	Ships, boats and floating structures
90	Optical, photographic, cinematographic, measuring, checking, precision, medical or surgical instruments and apparatus; parts and accessories thereof
91	Clocks and watches and parts thereof
92	Musical instruments; parts and accessories of such articles
93	Arms and ammunition; parts and accessories thereof
94	Furniture; bedding, mattresses, mattress supports, cushions and similar stuffed furnishings
95	Toys, games and sports requisites; parts and accessories thereof
96	Miscellaneous manufactured articles
97	Works of art, collectors' pieces, and antiques

Table 19: Product codes and corresponding names at the HS-2 level.

Confusion matrix for products at the HS-4 and HS-2 levels

Fig. 22 reports, for the HS-4 level of product aggregation, the confusion matrix built from the normalized GENEPY[9] values associated with the products for the year 2018, computed respectively based on the incidence matrix $\mathbf{M}^\top$ and the surrogate incidence matrix $(\hat{\mathbf{M}}^\top)^{(MC)}$, then discretized in 8 classes according to the percentiles of the respective GENEPY distributions. The class 1 corresponds to the lowest GENEPY

[9]Since the GENEPY and $\widehat{\text{GENEPY}}^{(MC)}$ span different ranges, they have been rescaled to the same interval $[0, 1]$.

values, whereas the class 8 corresponds to the highest GENEPY values. Notice that the whole procedure applied to countries was repeated here from scratch starting from the transpose of the **RCA** matrix, focusing in this case on the products. Very similar results ($\tau_k \simeq 0.96$) are obtained if instead, one computes the GENEPY for products using as input the transpose of the surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$ obtained in the application of MC to countries, described in the main text. In the figure, the true classes are the ones computed starting from the GENEPY applied to the incidence matrix $\mathbf{M}^{\top}$, whereas the predicted classes are the ones computed starting from the GENEPY applied to the surrogate incidence matrix $(\hat{\mathbf{M}}^{\top})^{(MC)}$.

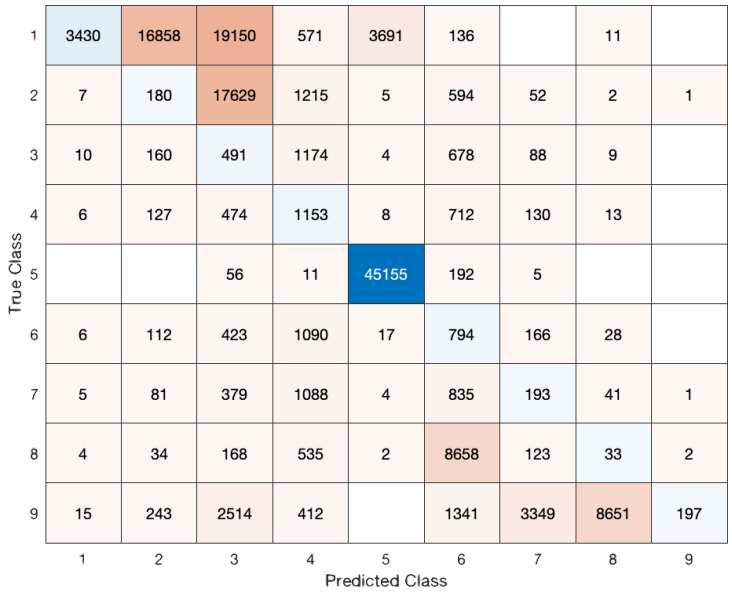


Figure 22: Confusion matrix built from the normalized 2018 GENEPY values associated with the products at the HS-4 level, computed respectively based on the matrices $\mathbf{M}^{\top}$ and $(\hat{\mathbf{M}}^{\top})^{(MC)}$, then discretized in 8 classes according to the percentiles of the respective GENEPY distributions.

As evidenced by Fig. 22, the GENEPY indices computed based on $\mathbf{M}^{\top}$ and $(\hat{\mathbf{M}}^{\top})^{(MC)}$ are quite similar, since the non-zero elements of the confusion matrix are concentrated above all near its main diagonal (i.e., the number of classification errors is quite small far from the main diagonal). Some outliers emerged. The deviations regarded mainly the elements belonging to higher categories. This is somewhat reasonable since they refer to more complex products, which might be more complex to classify.

In this case, the Kendall rank correlation coefficient between the GENEPY rankings computed based on $\mathbf{M}^\top$ and $(\hat{\mathbf{M}}^\top)^{(MC)}$ turned out to be $\tau_k \simeq 0.6$, with a p-value nearly equal to 0. Similarly, we built up a confusion matrix also for the HS-2 case. This is reported in Fig. 23.

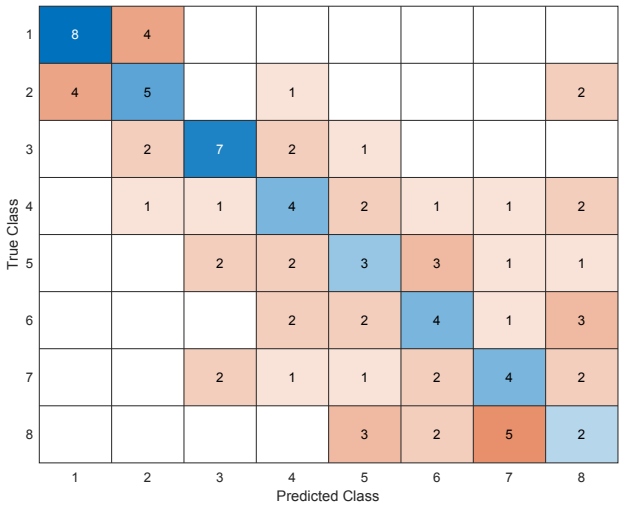


Figure 23: Confusion matrix built from the normalized 2018 GENEPY values associated with the products at the HS-2 level, computed respectively based on the matrices $\mathbf{M}^\top$ and $(\hat{\mathbf{M}}^\top)^{(MC)}$, then discretized in 8 classes according to the percentiles of the respective GENEPY distributions.

To corroborate the confusion matrix analysis at the product level, we discuss here if the false positive entries concentrate in the same products' category at the two aggregation levels (HS-2 and HS-4). Specifically, we included in Tab. 20 the classification of products by false positive rate. For the sake of clarity in the exposition, we report here only the top 11 products by HS classification:

fpr, **HS-2**	*fpr*, **HS-4**
Lead and articles thereof	Pharmaceutical products
Pearls, precious stones	Sugars and sugar confectionery
Fertilisers	Iron and steel
Copper and articles thereof	Lead and articles thereof
Zinc and articles thereof	Mineral fuels, mineral oils and products of their distillation; bituminous substances; mineral waxes
Tobacco and manufactured tobacco substitutes	Natural or cultured pearls, precious or semi-precious stones, precious metals
Mineral fuels, mineral oils (and related)	Oil seeds and oleaginous fruits; miscellaneous grains, seeds and fruit; industrial or medicinal plants
Other base metals; cermets; articles thereof	Vehicles other than railway or tramway rolling-stock, and parts and accessories thereof
Silk	Aluminium and articles thereof
Ores, slag and ash	Copper and articles thereof

Table 20: Top 11 products in terms of false positive rate at HS-2 and HS-4 levels of aggregation

For both HS-2 and HS-4 levels of aggregation, the analysis has been conducted by confronting the matrix of true values with the predicted ones obtained through 500 simulations (refer to the main text). This is to provide robustness to the results. Specifically, we recovered 500 predicted matrices (by applying the MC algorithm iteratively). Out of those 500 matrices, we recover a single predicted matrix by taking the 500 simulated prediction matrices (and discretizing it w.r.t the RCA being > 1). Then, the latter was used to construct the confusion matrix and compute the "by row" (by products) false-positive rate.

As we can see from Tab. 20 the false positives do not concentrate on the same products at the two aggregation levels, though some similarities emerge. Notice, in particular, how the HS-4 level seems to uncover the potentiality of different sectors with respect to HS-2 (e.g., Pharmaceutical and Vehicles). Since the sectors uncovered by the HS-4 level are associated with traditionally promising exporting sectors, they do not constitute a surprise. Instead, sectors such as precious minerals, fertilizers, aluminum, copper, iron, and steel reveal a high demand (overcoming the offer) of traditional technologies and raw materials in the world. The reasons might be multiple, but,

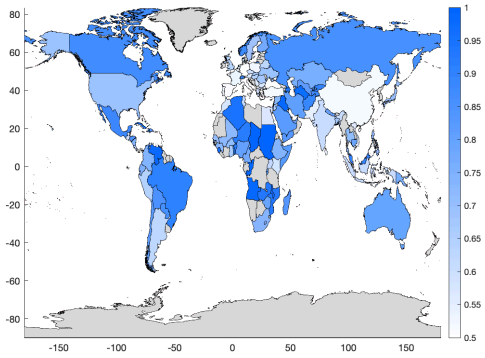
for sure, a key motivation is that such raw materials are mostly concentrated in less developed countries unable to go along with the world's needs of such goods.

Application of a variation of the analysis to countries, based on the entry-wise logarithm of the original RCA matrix for the year 2018, with products aggregated at the HS-4 level

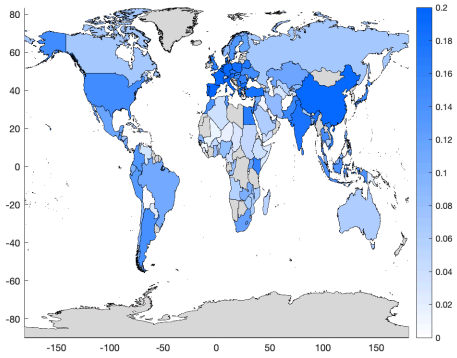
In the following, a variation of our analysis is applied to countries. In this variation, instead of discretizing the elements of the original **RCA** matrix, they are replaced by their natural logarithm[10]. Then, the rest of the proposed method is unchanged with respect to Section 2. In the main text, the discretization method has been preferred, because it generates values more symmetrically distributed around 0. The main motivations for the logarithm analysis are reported. First, the original data were continuous, and we did not want to lose the information about their continuity. Yet, in the main analysis, we categorized the original data. A further reason lies in the fact that matrix completion might work better when the elements have similar magnitude. Thus, we reduced the scale by taking the logarithm as a robustness check.

Fig. 24 reports, for this variation of analysis and for the product aggregation level HS-4, results similar to those obtained in Figs. 15(a) and 15(b) for the original analysis and for the product aggregation level HS-2.

[10]No issues arise when taking the natural logarithm, because no 0 or negative entries are present in that matrix. Moreover, entries originally equal to NaN are never included in the training, validation or test sets.



False negative rate fnr_c, reported proportionally to the shade of blue for the year 2018 and the product aggregation level HS-4, for the case in which the entry-wise natural logarithm of the original **RCA** matrix is employed by the proposed method. Countries colored in grey are not considered in the analysis.



False positive rate fpr_c, reported proportionally to the shade of blue for the year 2018 and the product aggregation level HS-4, for the case in which the entry-wise natural logarithm of the original **RCA** matrix is employed by the proposed method. Countries colored in grey are not considered in the analysis.

Figure 24: False negative and false positive rates for countries, obtained by the method reported in Step 5 of Section 2 for the year 2014, with products aggregated at the HS-2 level.

Similarly, Tab. 21 finds, for this variation of analysis and for the product aggregation level HS-4, results similar to those obtained in Tab. 18 for the original analysis and for the product aggregation level HS-2.

	GENEPY (τ_k)	**GENEPY** (p-**value**)
$fnr_{c,log,hs-4}$	-0.2569	0.0000
$fpr_{c,log,hs-4}$	0.5903	0.0000

Table 21: τ_k and relative p-values for the rankings produced using GENEPY against the ranking produced respectively by $fnr_{c,t}$ and $fpr_{c,t}$, for the product aggregation level HS-4, for the case in which the entry-wise natural logarithm of the original **RCA** matrix is employed by the original analysis.

Finally, also Fig. 25, which refers to the variation of analysis and to products aggregated at the HS-4 level, shows results similar to those obtained in Fig. 17(b), which refers to the original analysis and to products aggregated at the HS-2 level.

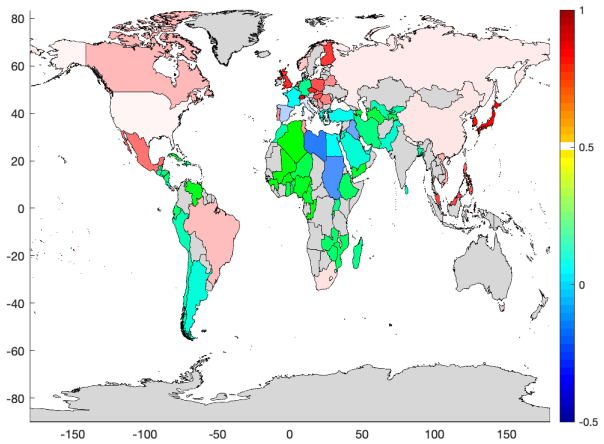


Figure 25: Difference between GENEPY and $\widehat{\text{GENEPY}}^{(MC)}$ for the year 2018, with products aggregated at the HS-4 level, for the case in which the entry-wise natural logarithm of the original **RCA** matrix is employed by the proposed method.

Conclusions

Machine learning techniques have advanced along with more sophisticated econometrics methodologies. Though the early practitioners of the former discipline were mainly statisticians and computer scientists, scholars (H. R. Varian, 2014a, Athey and Imbens, 2019) were able to transfer machine learning knowledge to other disciplines. In particular, machine learning has become more and more prominent in economics. As reported in Bargagli Stoffi (2020) *"The reasons behind the surge in the usage of machine learning are manifold and are connected to the major technological innovations introduced in the last decades. Advances in data storage, data transfer and data processing, and the availability of large data sources to be analyzed called for novel, more powerful, computational tools"*. However, such innovations come at a cost that can be either monetary or time spent to collect and manage data.

Machine learning techniques were, in the first instance, adopted for prediction within social sciences. Their advantage with respect to traditional econometrics methodologies was the ability of machine learning to uncover unexplored paths in data without the usual restrictions imposed by econometrics (Breiman, 2001). This does not mean that theories and assumptions are neglected in the machine learning era (Sejnowski, 2018).

On the contrary, as highlighted in Dominici, Bargagli-Stoffi, and Mealli (2020), machine learning alone cannot provide insights on the underlying assumptions of causal models. Nevertheless, a novel stream of machine learning is forming to take care also of causality issues (see, e.g., Athey, J. Tibshirani, and Wager, 2019, Bargagli-Stoffi, Niederreiter, and Riccaboni, 2021). However, the mentioned literature is still in an early stage and cannot deal with complex scenarios as the ones considered in Chapter 3 of the present Dissertation.

The main aim of this Dissertation is to extend the machine learning and econometrics frameworks through an application-oriented perspective. Indeed, the methods developed provide answers to relevant empirical questions. In the first two Chapters, we bridge several gaps in the field of applied machine learning to econometrics by providing a way to optimize the number of examples needed to conduct valid econometrics analyses in several scenarios. In the last two Chapters, we employ novel methodologies in machine learning and econometrics in order to provide innovative solutions to highly debated problems. Specifically, in Chapter 3 we introduce a methodological novelty to explore the relationship between market size and innovation, while in Chapter 4 we adopt MC – which was traditionally used in ML to impute unknown matrix entries – to build up a novel economic complexity index. Table 22 provides an overview of the main empirical contributions of the Dissertation.

6 Contributions

In the first Chapter, we introduced a novel machine learning methodology to investigate the optimal trade-off between the sample size and the precision of supervision in several econometrics contexts. The problem is fascinating in the era of big data, where the debate on their sunk costs is still vivid (Davenport and Dyché, 2013). Yet, big data may be costly not only in terms of money and time (Singh and Reddy, 2015) but also in terms of complexity (Tole et al., 2013). In particular, the latter comprehends difficulties in managing big data as well as in understanding them. Therefore, having the possibility of analyzing a smaller amount of data being sure to produce robust statistics could provide advantages in the two latter senses. The aim of Part 1 is to facilitate such theoretical effort through machine learning techniques by providing robust boundaries on the generalization error made when investigating "many but bad examples" or "few but good ones."

We considered three common scenarios in both macroeconometrics and microeconometrics applications. In all such cases, we assume a panel-data setting, i.e. N units observed in T time periods. Specifically, we studied a baseline model in which we

Table 22: The Table shows the contributions of the present Dissertation. {\footnotesize Note: the policy implications illustrated in the Table are just some of the various, possible implications connected to our empirical findings. The same holds true for further lines of research. Please refer to the related Chapters for more detailed discussions.}

	Chapter	Research Question	Findings	Policy Implications	Further Research
Part 1	Chapter 1	How should one optimally choose examples in a fixed effect unbalanced context?	"many but bad" examples provide a smaller upper bound on the conditional generalization error than "few but good" ones, whereas in other cases the opposite occurs	Optimal time and money spending on data collection in institutions	Extension to real world data; extension to non-linear models (e.g. Poisson FE).
	Chapter 2	How should one optimally choose examples in a fixed effect context with correlated errors?	the choice between 'many but bad' examples and 'few but goods' depends on the parameter α	Optimal time and money spending on data collection in institutions	Extension to unbalanced FEGLS; extension to causal inference and real world data.
Part 2	Chapter 3	Can recalls help investigating the relationship between market size and innovation?	The use of recalls successfully solves the reverse causality of market size and innovation revealing a positive causal relation between the two.	Targeted public help on firms; better funding allocation to innovate; M&A and partnerships of small bio-labs with Big Pharma to pursue innovation; renewed view of recalls in pharma.	Further analyses on product and firm levels; employ machine learning to predict cross-price elasticities of drugs; employ more up-to-date data in order to include also recalls of compounders and repackaging firms.
	Chapter 4	How can we measure economic complexity through machine learning?	We build up the innovative MONEY index to rank countries according to their economic complexity	Provide a recommender system for nations to build up economic relationships; better decision making based on economic complexity of countries; predictive power on future economic growth of nations.	Optimize MC algorithm; provide an analysis at the HS-6 level of aggregation.

eliminated the individual unobserved fixed effects. These effects are time-independent but individual-dependent. See, for instance, Burke and Sass (2013) for an example of how fixed effects can be used to control for individual effects (teachers and peers in their paper). In the last part of Chapter 1, we moved away from the baseline model and considered when individuals are not observed for the same amount of time. Formally, N units are now observed for a *maximum* time of T. Each unit is instead observed for an amount T_n of time. In this scenario, however, the measurements errors of observations associated with the same unit are mutually independent. In Chapter 2, we removed this hypothesis and considered the more general case in which the latter does not hold. Furthermore, we provided large-sample approximations with respect to the number of units and the number of observations. The results are tested numerically.
In both Chapters, we obtain similar quantitative results. These show that in the case of "decreasing returns of scale," "many but bad" examples are associated with a smaller conditional generalization error than "few but good" ones. The opposite occurs for "increasing returns of scale," whereas the "constant returns of scale" are intermediate.

While the first two Chapters fill the gaps left in the theoretical literature of machine learning and econometrics, the third and the fourth Chapters are more on the applicative side of the latter stream.
Chapter 3 investigated an example of when econometrics techniques provide more up-to-date, robust tools for a specific type of causality problem. The open debate on how and how much market size influences innovation requires techniques able to control at the same time for idiosyncratic and heterogeneity endogeneity and fixed effects in a counting data environment. This is, to the best of our knowledge, only possible with recently developed econometrics tools (W. Lin and Wooldridge, 2019). In dealing with the issue, we employed a novel instrument, namely recall, to solve for the reverse causality of market size and innovation. The results confirm the presence of the latter source of endogeneity and provide a robust point estimate of the relationship between market size and innovation, which turned out to be significant and positive. The analysis has been extended to product and firm levels of aggregation in

the Appendix. Further and novel insights on recalls have been discussed throughout the Dissertation.
Finally, Chapter 4 concludes with an application of machine learning to traditional economic issues, namely economic complexity. The topic of economic complexity has been indeed debated in the literature for long and for a long time (Keeley, 1988,Hidalgo and Hausmann, 2009,Simoes and Hidalgo, 2011 among others). This interest lies, above all, in the possibility of predicting the economic growth of countries in the future. As highlighted in the Introduction of the present Dissertation, machine learning provides, nowadays, novel tools to deal with predictions. In Chapter 4 we, particularly, exploited Matrix Completion (MC) in order to impute the so called Revealed Comparative Advantage of countries [11]. The overall idea to construct a novel index that innovates with respect to the elder ones, is that the worse the prediction of MC the more complex the country must be (otherwise MC would not have any difficulty in imputing the RCA of such a country). Being the first time that MC is employed to construct a complexity index, we believe that the novelty of MONEY lies on its capability to catch the complexity of countries by considering, at the same time, the complexity of the goods that they trade.

7 Outline of possible future research

Due to the scarcity in the literature concerning the first two Chapters, we invite authors to explore more complicated scenarios than the ones presented in the present Dissertation. The following list provides possible extensions:

- Application of he proposed methodologies to real world data. It would be very interesting to check that what computed in theory and simulated with algorithms is also valid with real world data. The difficulty lies on finding data aligning with the restrictions imposed throughout thee discussion;

- extension to causal inference [12]; (Shalit, Johansson, and Sontag, 2017)

[11] an index calculating the relative advantage or disadvantage of a certain country in a certain class of goods or services as evidenced by trade flows

[12] we remember here hat, in the context discussed in Chapters 1 and 2 the aim was to minimize a generalization error and hence the problem was mainly a prediction problem

- exploring the case of Instrumental Variables (IV) when classic OLS assumption fail and endogeneity intervenes;
- extending the analysis to more complex (i.e. having less and less restrictive assumptions) econometric models (unbalanced FEGLS, Poisson, Random Effects, Fixed Effect Poisson and many others)
- substitution of empirical bounds with bounds proved in literature.

As an example of application of the prospective extensions to a causal inference scenario to data analysis, one could jointly optimize the number of persons testing an upgraded version of a social media platform (e.g., one based on a new graphical interface, or containing new functionalities, such as the integration with other apps including item recommendations systems based on user's behavior) and the number of persons selected as controls, to assess the heterogeneous causal effects on the perceived quality of that upgrade. This optimization would be constrained by a budget constraint on the total supervision cost, and the possibility to enlarge (respectively, reduce) the two training set sizes above by reducing (respectively, increasing) the quality of each supervision – in case this was associated with a better estimate of the heterogeneous causal effects –, still respecting the given budget constraint.
Chapter 3 regards the application of recent econometric tools to explore the relationship between market size and innovation. Future research could employ more up-to-date data in order to also take into account in the analysis repackaging firms. Moreover, one can also employ double machine learning techniques to compute the own and cross price elasticties of prices of drugs. The latter information, by forecasting the co-movements of demand and price of medicines could be used as a further control in the study. However, this is not an easy task since computing cross-elasticities within a panel-data model may require a huge computational time. Computing cross-elasticities, moreover, is a simultaneous equation problem which can have no closed form. Sophisticated and novel tools are, thus, required.
Finally, Chapter 4 could be easily improved by optimizing the actual algorithm used to perform MC and extend the actual number of iterations. The optimized MC algorithm could be also employed to extend the analysis at the HS-6 level of aggregation to provide better recommendations to countries. Future research should also apply

MONEY index in economics context and check its provided ranking.

Appendix A

Appendix

The Appendices for Chapter 1 and Chapter 2 come from Gnecco, Nutarelli, and Selvi (2020) and Gnecco, Nutarelli, and Selvi (2021).

1 Appendix for Chapter 1

The following Appendix includes the proofs of formulas in Chapter 1. This Appendix does not include proofs for the baseline model since it was only for explanatory purposes (see Gnecco and Nutarelli, 2019b for further details).
First, we expand the conditional generalization error (1.32) as follows:

$$\begin{aligned}
&\mathbb{E}\left\{\left(\hat{\eta}_{i,FE}+\hat{\boldsymbol{\beta}}'_{FE}\boldsymbol{x}_i^{test}-\eta_i-\boldsymbol{\beta}'\boldsymbol{x}_i^{test}\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
=&\mathbb{E}\left\{\left((\hat{\eta}_{i,FE}-\eta_i)+\left(\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{test}\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
=&\mathbb{E}\left\{(\hat{\eta}_{i,FE}-\eta_i)^2|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&+\mathbb{E}\left\{\left(\left(\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{test}\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&+2\mathbb{E}\left\{(\hat{\eta}_{i,FE}-\eta_i)\left(\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{test}\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}. \qquad \text{(A.1)}
\end{aligned}$$

Exploiting the conditional unbiasedness of $\hat{\eta}_{i,FE}$, and the expressions (2.1) of $y_{n,t}$, (**??**) of $\tilde{y}_{n,t}$, and (2.12) of $\hat{\eta}_{i,FE}$ (with the index n replaced by the index i), one gets

$$\mathbb{E}\left\{(\hat{\eta}_{i,FE}-\eta_i)^2|\{\boldsymbol{X}_n\}_{n=1}^N\right\}$$

$$
\begin{aligned}
=&\mathbb{E}\Bigg\{\left(\frac{1}{T_i}\left(\sum_{t=1}^{T_i}\left(\eta_i+\boldsymbol{\beta}'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}-\hat{\boldsymbol{\beta}}'_{FE}\boldsymbol{x}_{i,t}\right)\right)-\eta_i\right)^2 \\
&\qquad\Big|\{\boldsymbol{X}_n\}_{n=1}^N\Bigg\} \\
=&\mathbb{E}\left\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{FE}\right)'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}\right)\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}. \qquad \text{(A.2)}
\end{aligned}
$$

It follows from Equation (A.2) that Equation (A.1) can be re-written as

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{FE}\right)'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}\right)\right)^2\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\} \\
&+\mathbb{E}\left\{\left(\boldsymbol{x}_i^{test}\right)'\left(\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{test}\Big|\{\boldsymbol{X}_n\}_{n=1}^N\right\} \\
&+2\mathbb{E}\Bigg\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{FE}\right)'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}\right)\right) \\
&\qquad\left(\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{test}\Big|\{\boldsymbol{X}_n\}_{n=1}^N\Bigg\}. \qquad \text{(A.3)}
\end{aligned}
$$

Using the expression (2.9) of $\hat{\boldsymbol{\beta}}_{FE}$, and Equation (A.1), one can simplify the term $\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}$ above as follows:

$$
\begin{aligned}
&\hat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta} \\
=&\left(\sum_{n=1}^N \boldsymbol{X}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\left(\sum_{n=1}^N \boldsymbol{X}_n'\boldsymbol{Q}_n\tilde{\boldsymbol{y}}_n\right)-\boldsymbol{\beta} \\
=&\left(\sum_{n=1}^N \boldsymbol{X}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\left(\sum_{n=1}^N \boldsymbol{X}_n'\boldsymbol{Q}_n\left(\boldsymbol{\eta}_n+\boldsymbol{X}_n\boldsymbol{\beta}+\boldsymbol{\varepsilon}_n\right)\right)-\boldsymbol{\beta} \\
=&\left(\sum_{n=1}^N \boldsymbol{X}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\left(\sum_{n=1}^N \boldsymbol{X}_n'\boldsymbol{Q}_n\boldsymbol{\varepsilon}_n\right). \qquad \text{(A.4)}
\end{aligned}
$$

Then, Equation (A.3) becomes

$$
\mathbb{E}\Bigg\{\left(\frac{\mathbf{1}'_{T_i}}{T_i}\left(-\boldsymbol{X}_i\left(\sum_{n=1}^N \boldsymbol{X}_n'\boldsymbol{Q}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\right.\right.
$$

$$
\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) + \boldsymbol{\varepsilon}_i\Bigg)\Bigg)^2 \Big| \{\boldsymbol{X}_n\}_{n=1}^N \Bigg\}
$$

$$
+ \mathbb{E}\Bigg\{ \left(\boldsymbol{x}_i^{test}\right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right)
$$

$$
\left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right)\right)' \boldsymbol{x}_i^{test}
$$

$$
\Big| \{\boldsymbol{X}_n\}_{n=1}^N \Bigg\}
$$

$$
+ 2\mathbb{E}\Bigg\{ \Bigg(\frac{\mathbf{1}_{T_i}'}{T_i} \Bigg(- \boldsymbol{X}_i \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}
$$

$$
\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) + \boldsymbol{\varepsilon}_i \Bigg)\Bigg)
$$

$$
\left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right)\right)' \boldsymbol{x}_i^{test}
$$

$$
\Big| \{\boldsymbol{X}_n\}_{n=1}^N \Bigg\}. \tag{A.5}
$$

Expanding the square in the first term in the expression above, and splitting its last term in two parts, one obtains the following expression for Equation (A.5):

$$
\mathbb{E}\Bigg\{ \frac{\mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right)
$$

$$
\left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)'
$$

$$
\boldsymbol{X}_i' \mathbf{1}_{T_i} \Big| \{\boldsymbol{X}_n\}_{n=1}^N \Bigg\}
$$

$$
+ \mathbb{E}\left\{ \frac{\mathbf{1}_{T_i}' \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \mathbf{1}_{T_i}}{T_i^2} \Big| \{\boldsymbol{X}_n\}_{n=1}^N \right\}
$$

$$
- 2\mathbb{E}\Bigg\{ \frac{\mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}
$$

$$
\begin{aligned}
&\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \boldsymbol{\varepsilon}_i' \mathbf{1}_{T_i} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
&+ \mathbb{E}\Bigg\{ \left(\boldsymbol{x}_i^{test}\right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
&- 2\mathbb{E}\Bigg\{ \frac{\mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
&+ 2\mathbb{E}\Bigg\{ \frac{\mathbf{1}_{T_i}' \boldsymbol{\varepsilon}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\}.
\end{aligned}
\tag{A.6}
$$

In order to simplify the various terms contained in Equation (A.6), one observes that, due to Equations (A.7), (A.8), and (1.36), one gets

$$
\begin{aligned}
&\mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \\
&\quad \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
=&\mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
=&\sigma^2 \mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
=&\sigma^2 \mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
=&\sigma^2 \mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\}, \qquad \text{(A.7)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \boldsymbol{\varepsilon}_i' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
=&\mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \boldsymbol{Q}_i' \boldsymbol{Q}_i \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\} \\
=&\sigma^2 \mathbb{E}\Bigg\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \boldsymbol{Q}_i' \boldsymbol{Q}_i \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \Bigg\}. \qquad \text{(A.8)}
\end{aligned}
$$

Then, by an application of the two equations just derived above, one obtains the following equivalent expression for Equation (A.6):

$$
\begin{aligned}
&\frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{X}_i' \mathbf{1}_{T_i} \\
&+ \frac{\sigma^2}{T_i} \\
&- 2 \frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{X}_i' \boldsymbol{Q}_i' \boldsymbol{Q}_i \mathbf{1}_{T_i} \\
&+ \mathbb{E} \left\{ \sigma^2 \left(\boldsymbol{x}_i^{test} \right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \right\} \\
&- 2 \mathbb{E} \left\{ \frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \right\} \\
&+ 2 \mathbb{E} \left\{ \frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{Q}_i' \boldsymbol{Q}_i \boldsymbol{X}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \right\},
\end{aligned}
\tag{A.9}
$$

where, in some cases, the conditional expectations of deterministic matrices (and of random matrices, like $\boldsymbol{X}_i$, that become known once the set of conditioning matrices $\{\boldsymbol{X}_n\}_{n=1}^{N}$ has been fixed) have been replaced by the matrices themselves. Finally, exploiting Equation (1.38), one can get rid of the third and sixth terms in Equation (A.9), which then becomes

$$
\begin{aligned}
&\frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{X}_i' \mathbf{1}_{T_i} \\
&+ \frac{\sigma^2}{T_i} \\
&+ \mathbb{E} \left\{ \sigma^2 \left(\boldsymbol{x}_i^{test} \right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \right\} \\
&- 2 \mathbb{E} \left\{ \frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \right)^{-1} \boldsymbol{x}_i^{test} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N} \right\},
\end{aligned}
\tag{A.10}
$$

which is Equation (2.17).

1.1 Proof of Equation (1.47)

The first inequality

$$\begin{aligned}&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\\ \leq&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\lambda_{\max}(\boldsymbol{A}_N^{-1})\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\end{aligned} \tag{A.11}$$

in Equation (1.47) is obtained by exploiting the definition of induced l_2-matrix norm, i.e.,

$$\|\boldsymbol{A}_N^{-\frac{1}{2}}\|_2=\sup_{\boldsymbol{x}\in\mathbb{R}^p,\|\boldsymbol{x}\|_2\neq 0}\frac{\|\boldsymbol{A}_N^{-\frac{1}{2}}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}\,, \tag{A.12}$$

and the fact that, being $\boldsymbol{A}_N^{-\frac{1}{2}}$ symmetric, one has

$$\|\boldsymbol{A}_N^{-\frac{1}{2}}\|_2^2=\lambda_{\max}^2(\boldsymbol{A}_N^{-\frac{1}{2}})=\lambda_{\max}(\boldsymbol{A}_N^{-1})\,. \tag{A.13}$$

Then, the equality

$$\begin{aligned}&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\lambda_{\max}(\boldsymbol{A}_N^{-1})\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\\ =&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\frac{1}{\lambda_{\min}(\boldsymbol{A}_N)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{test}\right)\right\|_2^2\right\}\right)\,.\end{aligned} \tag{A.14}$$

follows from the relationship $\lambda_{\min}(\boldsymbol{A}_N)=\frac{1}{\lambda_{\max}(\boldsymbol{A}_N^{-1})}$.

Finally, the last inequality in Equation (1.47) is obtained by exploiting Weyl's inequalities Bhatia, 1997, Theorem III.2.1 for the eigenvalues of the sum of symmetric matrices, as detailed in the following remark.

Remark 7. *Given any pair of symmetric matrices $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{s\times s}$, let their eigenvalues and those of $\boldsymbol{C}:=\boldsymbol{A}+\boldsymbol{B}$ be ordered nondecreasingly (with possible repetitions in case of multiplicity larger than 1) as*

$$\begin{aligned}&\lambda_1(\boldsymbol{A})\leq\lambda_2(\boldsymbol{A})\leq\ldots\leq\lambda_k(\boldsymbol{A})\leq\ldots\leq\lambda_s(\boldsymbol{A})\,,\\ &\lambda_1(\boldsymbol{B})\leq\lambda_2(\boldsymbol{B})\leq\ldots\leq\lambda_k(\boldsymbol{B})\leq\ldots\leq\lambda_s(\boldsymbol{B})\,,\\ &\lambda_1(\boldsymbol{C})\leq\lambda_2(\boldsymbol{C})\leq\ldots\leq\lambda_k(\boldsymbol{C})\leq\ldots\leq\lambda_s(\boldsymbol{C})\,.\end{aligned} \tag{A.15}$$

Then, Weyl's inequalities, in their simplest form, state that, for every $k = 1, \ldots, s$, *one has*

$$\lambda_k(\boldsymbol{A}) + \lambda_1(\boldsymbol{B}) \leq \lambda_k(\boldsymbol{C}) \leq \lambda_k(\boldsymbol{A}) + \lambda_s(\boldsymbol{B}) \,. \tag{A.16}$$

Hence, $\lambda_{\min}(\boldsymbol{C}) \geq \lambda_{\min}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B})$. *Similarly, for any* $\mu_1, \mu_2 \geq 0$, *when* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are also positive semi-definite (as in the case of the matrices* $\boldsymbol{A}$ *defined in Equation (1.40)), one gets*

$$\lambda_{\min}(\mu_1 \boldsymbol{A} + \mu_2 \boldsymbol{B}) \geq \mu_1 \lambda_{\min}(\boldsymbol{A}) + \mu_2 \lambda_{\min}(\boldsymbol{B}) \,. \tag{A.17}$$

Finally, Equation (A.17) extends directly to the case of a weighted summation (with non-negative weights) of symmetric and positive semi-definite matrices, proving the last inequality in Equation (1.47).

2 Appendix for Chapter 2

Appendix 1: proofs of Eqs. (2.17), (2.18), and (2.19)

To simplify the notation, here and in the next appendices, Q_T will be often replaced by the shorthand Q. Also the dependence of Ψ and Φ and other matrices/vectors on T will be typically omitted in the notation, apart from a few cases (e.g., in part of Appendix 2).

Proof of Eq. (2.17)

For $n = 1, \ldots, N$, let $\underline{\eta}_n \in \mathbb{R}^T$ be the column vector whose elements are all equal to η_n. Using the expressions (2.1), (2.2), (2.6), and (2.9) respectively of $y_{n,t}$, $z_{n,t}$, $\underline{\ddot{z}}_n$, and $\underline{\hat{\beta}}_{FEGLS}$, and

$$Q\underline{\eta}_n = \underline{0}_T \,, \tag{A.1}$$

one can write the term $\underline{\hat{\beta}}_{FEGLS} - \underline{\beta}$ as follows:

$$\begin{aligned} \underline{\hat{\beta}}_{FEGLS} - \underline{\beta} &= \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{X}_n \right)^{-1} \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \underline{\ddot{z}}_n \right) - \underline{\beta} \\ &= \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{X}_n \right)^{-1} \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ Q \left(\underline{\eta}_n + X_n \underline{\beta} + \underline{\varepsilon}_n \right) \right) - \underline{\beta} \end{aligned}$$

$$= \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{X}_n\right)^{-1} \sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{\underline{\varepsilon}}_n . \qquad (A.2)$$

In the following, to simplify the notation, we set $B := \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{X}_n\right)^{-1}$ and $\underline{b} := \sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{\underline{\varepsilon}}_n$.

Now, we expand the conditional generalization error (2.16) as follows:

$$\begin{aligned} R_i\left(\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right) &= \mathbb{E}\left\{\left(\hat{\eta}_{i,FEGLS} + \hat{\underline{\beta}}'_{FEGLS}\underline{x}_i^{test} - \eta_i - \underline{\beta}'\underline{x}_i^{test}\right)^2 \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\ &= \mathbb{E}\left\{(\hat{\eta}_{i,FEGLS} - \eta_i)^2 | \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} + \mathbb{E}\left\{\left(\left(\hat{\underline{\beta}}_{FEGLS} - \underline{\beta}\right)' \underline{x}_i^{test}\right)^2 \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\ &+ 2\mathbb{E}\left\{(\hat{\eta}_{i,FEGLS} - \eta_i)\left(\hat{\underline{\beta}}_{FEGLS} - \underline{\beta}\right)' \underline{x}_i^{test} \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} . \qquad (A.3) \end{aligned}$$

Exploiting the conditional unbiasedness of $\hat{\eta}_{i,FEGLS}$, and the expressions (2.1) of $y_{n,t}$, (2.2) of $z_{n,t}$, and (2.12) of $\hat{\eta}_{i,FEGLS}$ (with the index n replaced by the index i), one gets

$$\begin{aligned} &\mathbb{E}\left\{(\hat{\eta}_{i,FEGLS} - \eta_i)^2 | \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\ = \; &\mathbb{E}\left\{\left(\frac{1}{T}\sum_{t=1}^{T}\left(\left(\underline{\beta} - \hat{\underline{\beta}}_{FEGLS}\right)' \underline{x}_{i,t} + \varepsilon_{i,t}\right)\right)^2 \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} . \qquad (A.4) \end{aligned}$$

Then, taking into account (A.2) and (A.4), one gets

$$\begin{aligned} (A.3) \;=\; &\mathbb{E}\left\{\left(\frac{1}{T}\underline{1}_T'(-X_i B\underline{b} + \underline{\varepsilon}_i)\right)^2 \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} + \mathbb{E}\left\{\left(\underline{x}_i^{test}\right)' B\underline{b}(B\underline{b})' \underline{x}_i^{test} \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\ &+ 2\mathbb{E}\left\{\left(\frac{1}{T}\underline{1}_T'(-X_i B\underline{b} + \underline{\varepsilon}_i)\right)(B\underline{b})' \underline{x}_i^{test} \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} . \qquad (A.5) \end{aligned}$$

Expanding the square in the first term in the expression above, and splitting its last term in two parts, one obtains

$$\begin{aligned} (A.5) \;=\; &\mathbb{E}\left\{\frac{1}{T^2}\underline{1}_T' X_i B\underline{b}\underline{b}' B' X_i' \underline{1}_T \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} + \mathbb{E}\left\{\frac{1}{T^2}\underline{1}_T' \underline{\varepsilon}_i \underline{\varepsilon}_i' \underline{1}_T \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\ &- 2\mathbb{E}\left\{\frac{1}{T^2}\underline{1}_T' X_i B\underline{b}\underline{\varepsilon}_i' \underline{1}_T \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} + \mathbb{E}\left\{\left(\underline{x}_i^{test}\right)' B\underline{b}\underline{b}' B' \underline{x}_i^{test} \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\ &- 2\mathbb{E}\left\{\frac{1}{T}\underline{1}_T' X_i B\underline{b}\underline{b}' B' \underline{x}_i^{test} \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} + 2\mathbb{E}\left\{\frac{1}{T}\underline{1}_T' \underline{\varepsilon}_i \underline{b}' B' \underline{x}_i^{test} \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} . \\ &\qquad (A.6) \end{aligned}$$

In order to simplify the expressions above, one exploits the following properties:

$$\mathbb{E}\left\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}'_m\right\} = \underline{0}_{T\times T} \tag{A.7}$$

for $n \neq m$ (being $\underline{0}_{T\times T} \in \mathbb{R}^{T\times T}$ the matrix whose elements are all equal to 0), and

$$\mathbb{E}\left\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}'_n\right\} = \Omega = \sigma^2\Phi\,. \tag{A.8}$$

Then, expanding $\underline{b}$ and exploiting also the facts that $\Omega^+\Omega\Omega^+ = \Omega$, $B = B'$, the matrix Ω^+ is symmetric and deterministic, and all the $\ddot{X}_n$ are known once all the $\underline{x}_{n,t}$ are given, one gets

$$\begin{aligned}
\mathbb{E}\left\{B\underline{b}\underline{b}'B' \middle| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} &= \mathbb{E}\left\{B\left(\sum_{n=1}^{N} \ddot{X}'_n\Omega^+\ddot{\underline{\varepsilon}}_n \sum_{m=1}^{N} \ddot{\underline{\varepsilon}}'_m\Omega^+\ddot{X}_m\right)B' \middle| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\
&= B\left(\sum_{n=1}^{N} \ddot{X}'_n\Omega^+\Omega\Omega^+\ddot{X}_n\right)B' \\
&= BB^{-1}B' \\
&= \sigma^2\left(\sum_{n=1}^{N} \ddot{X}'_n\Phi^+\ddot{X}_n\right)^{-1}.
\end{aligned} \tag{A.9}$$

Finally, inserting (A.9) in (A.6), expanding the expressions of B and $\underline{b}$, and recalling that $\mathbb{E}\{\underline{\varepsilon}_i\underline{\varepsilon}'_i\} = \sigma^2\Psi$, one obtains Eq. (2.17).

Proof of Eq. (2.18)

The expression $\underline{1}'_T\Psi\underline{1}_T$ in Eq. (2.18) is the summation of all the elements of the matrix Ψ. Now, the element $\rho^0 = 1$ appears in that summation T times, whereas the generic element ρ^t (for $t = 1,\ldots,T-1$) appears $2(T-t)$ times. Hence,

$$\underline{1}'_T\Psi\underline{1}_T = \left(T + \sum_{t=1}^{T-1} 2(T-t)\rho^t\right) = \left(T + 2T\sum_{t=1}^{T-1}\rho^t - 2\sum_{t=1}^{T-1} t\rho^t\right). \tag{A.10}$$

Then, Eq. (2.18) is obtained from (A.10) by exploiting the following well-known expressions for the partial sums of the geometric series, and of its derivative:

$$\sum_{t=1}^{T-1}\rho^t = \frac{1-\rho^T}{1-\rho} - 1\,, \tag{A.11}$$

and

$$\sum_{t=1}^{T-1} t\rho^t = \rho\frac{d\left(\sum_{t=1}^{T-1}\rho^t\right)}{d\rho} = \frac{\rho}{1-\rho}(-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \ldots + 1)\,, \tag{A.12}$$

where the right-hand side in Eq. (A.12) has been obtained by simplifying common

factors in the numerator and the denominator.

Proof of Eq. (2.19)

We compute $Q\Psi\underline{1}_T$, as follows:

$$\begin{aligned} Q\Psi\underline{1}_T &= \left(I_T - \frac{1}{T}\underline{1}_T\underline{1}_T{}'\right)\Psi\underline{1}_T \\ &= \Psi\underline{1}_T - \frac{1}{T}(\underline{1}_T'\Psi\underline{1}_T)\underline{1}_T \\ &= \Psi\underline{1}_T - \left[1 + 2\left(\frac{1-\rho^T}{1-\rho} - 1\right) - \frac{2\rho}{T(1-\rho)}\Big(-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1\Big)\right]\underline{1}_T, \end{aligned} \tag{A.13}$$

where the expression above of $(\underline{1}_T'\Psi\underline{1}_T)\underline{1}_T$ comes from Eq. (2.18). Finally, Eq. (2.19) is obtained from Eq. (A.13) by expanding the elements of $\Psi\underline{1}_T$, then simplifying the resulting expressions.

Appendix 2: proofs of the probability limits in Section 2

In the following, Eqs. (2.20) and (2.21) are derived under the common assumption that, for the unit i, the $\underline{x}_{i,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4. To derive Eq. (2.23), one makes the similar assumption that, for each fixed unit n, the $\underline{x}_{n,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4, together with the additional assumption $\lim_{T\to\infty} \|\Phi^+ - Q\Psi^{-1}Q'\|_2 = 0$ reported in Eq. (2.22). The validity of this last assumption is discussed extensively at the end of this appendix. Eqs. (2.20), (2.21), and (2.23) could be derived under more general conditions, but such possible extension is out of the scope of the paper.

Proof of Eq. (2.20)

Eq. (2.20) simply replaces the empirical average of the transposes of the $\underline{x}_{n,t}$ (which is $\frac{1}{T}\underline{1}_T'X_i$) with their common expected value $\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)'$, and follows from Chebyschev's weak law of large numbers (Ruud et al., 2000, Section 13.4.2).

Proof of Eq. (2.21)

In order to prove Eq. (2.21), it is convenient to introduce (recalling the definition of $\underline{u}_T$ provided in Eq. (2.19)) the vector

$$\underline{v}_T := Q'\Phi^+ Q\Psi \underline{1}_T = Q'\Phi^+ \underline{u}_T\,, \tag{A.14}$$

since the argument of the probability limit in Eq. (2.21) can be written as follows:

$$\frac{1}{T}\ddot{X}_i'\Phi^+ Q\Psi\underline{1}_T = \frac{1}{T}X_i'Q'\Phi^+ Q\Psi\underline{1}_T = \frac{1}{T}X_i'\underline{v}_T\,. \tag{A.15}$$

In other words, the T elements of each row of X_i' are summed with (different and deterministic) weights $v_{T,t}$ (the components of $\underline{v}_T$), for $t=1,\ldots,T$, then their weighted sum is divided by T. This suggests the application of a suitable form of the law of large numbers, which holds in this case: specifically, the one provided in Bai, P. E. Cheng, and Zhang (1997, Theorem 2.1)). In view of the next application of that theorem, first we investigate the following properties of the various terms involved in Eqs. (A.14) and (A.15).

i) *The Euclidean norm of the vector $\underline{u}_T$ is bounded from above as follows, for $K_u>0$ independent from T:*

$$\|\underline{u}_T\|_2 \leq K_u\sqrt{T}\,. \tag{A.16}$$

This follows from the fact that the absolute values of all the components of the vector $\underline{u}_T$, whose expression is reported in Eq. (2.19), are bounded from above by a sufficiently large $K_u>0$, which is independent from T.

ii) *All the eigenvalues of the matrix Ψ belong to the interval $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset (0,+\infty)$.*

This result follows by observing that Ψ is a symmetric Toeplitz matrix[1] (Gray, 2006). Then, by Gray (2006, Lemma 4.1), all the eigenvalues of Ψ belong to

[1] We recall that a matrix $M\in\mathbb{R}^{T\times T}$ is a symmetric Toeplitz matrix if it has the form

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \cdots & m_{T-2} & m_{T-1} \\ m_1 & m_0 & m_1 & m_2 & \cdots & m_{T-2} \\ m_2 & m_1 & m_0 & m_1 & \cdots & m_{T-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{T-1} & m_{T-2} & \cdots & m_2 & m_1 & m_0 \end{bmatrix}, \tag{A.17}$$

where $m_0, m_1, \ldots, m_{T-1}\in\mathbb{R}$.

the interval $[m_f, M_f]$, where m_f and M_f are respectively the minimum and the maximum of the function

$$f(\lambda) := \sum_{k=-\infty}^{+\infty} \rho^{|k|} e^{\iota k \lambda} \tag{A.18}$$

on the interval $[0, 2\pi]$, and ι is the imaginary unit. By inverting the Fourier series (A.18) as in Gilgen (2006, Eqs. (7.77)-(7.79)), one gets

$$f(\lambda) = \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\lambda)}, \tag{A.19}$$

from which one gets $m_f = \frac{1-\rho^2}{1+\rho^2+2\rho} > 0$ and $M_f = \frac{1-\rho^2}{1+\rho^2-2\rho} < +\infty$ since $\rho \in (-1, 1)$, which concludes the proof of item ii).

iii) *The matrix* Φ *has* 0 *as eigenvalue with multiplicity* 1*, and associated eigenvector* $\underline{1}_T$.

This result follows from the characterization of the eigenvalues of a symmetric matrix $M \in \mathbb{R}^{T \times T}$ as the stationary values of its Rayleigh quotient $\frac{\underline{x}' M \underline{x}}{\underline{x}' \underline{x}}$ (with $\underline{x} \in \mathbb{R}^T$ and $\underline{x} \neq \underline{0}_T$) (Parlett, 1998, Chapter 1), the invertibility of the matrix Ψ, and the fact that Q has eigenvalue 0 with multiplicity 1, and associated eigenvector $\underline{1}_T$ [2]. Hence, for $\underline{x} \neq \underline{0}_T$, $\frac{\underline{x}' M \underline{x}}{\underline{x}' \underline{x}} = 0$ if and only $\underline{x}$ is proportional to $\underline{1}_T$.

iv) *All the other eigenvalues of* Φ *belong to the interval* $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset (0, +\infty)$.

This follows again from the characterization of the eigenvalues of a symmetric matrix as the stationary values of its Rayleigh quotient, and also from Courant-Fisher's maxmin theorem (Parlett, 1998, Theorem 10.2.1) and from item ii). Indeed, by ordering the (real) eigenvalues of Ψ and Φ respectively as $\lambda_1(\Psi) \leq \lambda_2(\Psi) \leq \ldots \lambda_T(\Psi)$ and $\lambda_1(\Phi) \leq \lambda_2(\Phi) \leq \ldots \lambda_T(\Phi)$, and recalling that $\underline{x}' \Phi \underline{x} = \underline{x}' Q \Psi Q' \underline{x} = \underline{x}' \Psi \underline{x}$ for any $\underline{x} \in \mathbb{R}^T$ orthogonal to $\underline{1}_T$, one gets

$$\lambda_1(\Psi) = \min_{\underline{x} \in \mathbb{R}^T, \underline{x} \neq \underline{0}_T} \frac{\underline{x}' \Psi \underline{x}}{\underline{x}' \underline{x}} \leq \min_{\underline{x} \in \mathbb{R}^T, \underline{x} \neq \underline{0}_T, \underline{x} \perp \underline{1}_T} \frac{\underline{x}' \Psi \underline{x}}{\underline{x}' \underline{x}} = \min_{\underline{x} \in \mathbb{R}^T, \underline{x} \neq \underline{0}_T, \underline{x} \perp \underline{1}_T} \frac{\underline{x}' \Phi \underline{x}}{\underline{x}' \underline{x}} = \lambda_2(\Phi) \tag{A.20}$$

from Courant-Fisher's maxmin theorem, whereas

$$\lambda_T(\Psi) = \max_{\underline{x} \in \mathbb{R}^T, \underline{x} \neq \underline{0}_T} \frac{\underline{x}' \Psi \underline{x}}{\underline{x}' \underline{x}} \geq \max_{\underline{x} \in \mathbb{R}^T, \underline{x} \neq \underline{0}_T} \frac{\underline{x}' \Phi \underline{x}}{\underline{x}' \underline{x}} = \lambda_T(\Phi) \tag{A.21}$$

is obtained by expressing $Q' \underline{x}$ by using a basis of orthonormal eigenvectors $\underline{e}_t$

[2] remember that $\Phi = Q_T \Psi Q_T'$

of Ψ, associated with the respective eigenvalues $\lambda_t(\Psi)$, $t=1,\ldots,T$. Indeed, for some coefficients α_t, $t=1,\ldots,T$ (depending on $\underline{x}$), one has

$$Q'\underline{x} = \sum_{t=1}^{T} \alpha_t \underline{e}_t \,, \tag{A.22}$$

hence

$$\underline{x}'\Phi\underline{x} = \underline{x}'Q\Psi Q'\underline{x} = \sum_{t=1}^{T} \lambda_t(\Psi)\alpha_t^2 \,, \tag{A.23}$$

whereas

$$\underline{x}'\underline{x} \geq \underline{x}'QQ'\underline{x} = \sum_{t=1}^{T} \alpha_t^2 \,. \tag{A.24}$$

Then, one gets

$$\frac{\underline{x}'\Phi\underline{x}}{\underline{x}'\underline{x}} \leq \lambda_T(\Psi) \tag{A.25}$$

for any $\underline{x} \neq \underline{0}_T$.

v) *All the non-zero eigenvalues of the matrix Φ^+ belong to the interval $\left[\frac{1+\rho^2-2\rho}{1-\rho^2}, \frac{1+\rho^2+2\rho}{1-\rho^2}\right] \subset (0,+\infty)$.*

This follows from items iii) and iv) and the relation between the singular value decomposition of a symmetric positive semi-definite matrix and the singular value decomposition of its Moore-Penrose pseudoinverse, which has been reported in footnote 3.

vi) *The Euclidean norm of the vector $\underline{v}_T$ is bounded from above as follows, for $K_v > 0$ independent from T:*

$$\|\underline{v}_T\|_2 \leq K_v\sqrt{T} \,. \tag{A.26}$$

This is obtained by combining the definition of $\underline{v}_T$ provided in Eq. (A.14) with items i) and v), and the fact that the eigenvalue with maximus modulus of $Q = Q'$ is 1. A possible expression for K_v is $K_v = \frac{1+\rho^2+2\rho}{1-\rho^2}K_u$.

vii) *The following holds:*

$$\limsup_{T\to\infty} \frac{1}{\sqrt{T}} \|\underline{v}_T\|_2 \leq K_v < +\infty \,. \tag{A.27}$$

This is obtained immediately from item vi).

To conclude the proof of Eq. (2.21), we first consider the case in which, for the

unit i, all the $\underline{x}_{i,t}$ have mean $\underline{0}_p$. Later, this additional assumption is removed.

viii) *Proof of Eq. (2.21) when all the $\underline{x}_{i,t}$ have mean $\underline{0}_p$.*

Item vii) and the fact that all the elements of each row of X_i' have 0 mean, are independent, identically distributed, and their moments up to the order 4 are finite allow one to apply Bai, P. E. Cheng, and Zhang, 1997, Theorem 2.1, getting the following result, where, for $r = 1, \ldots, p$, $(X_i'\underline{v}_T)_r$ denotes the r-th component of $X_i'\underline{v}_T$:

$$\limsup_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T^{\frac{3}{4}}(\log T)^{\frac{1}{4}}} = 0 \text{ almost surely}^3 . \tag{A.28}$$

This, combined with the inequalities

$$0 \leq \liminf_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T} \leq \limsup_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T} \leq \limsup_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T^{\frac{3}{4}}(\log T)^{\frac{1}{4}}} \tag{A.29}$$

and the fact that almost sure convergence implies convergence in probability Rao et al., 1973, shows that, for all $r = 1, \ldots, p$, one has

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T}(X_i'\underline{v}_T)_r = 0 . \tag{A.30}$$

To conclude, one gets Eq. (2.21) from Eqs. (A.15) and (A.30), by exploiting the fact that, for a sequence of random matrices with fixed dimension, element-wise convergence in probability implies convergence in probability of the whole sequence Herriges, 2011.

ix) *Proof of Eq. (2.21) when all the $\underline{x}_{i,t}$ have the same mean $\underline{m} \in \mathbb{R}^p$.*

We set $\bar{\underline{x}}_{i,t} := \underline{x}_{i,t} - \underline{m}$, in such a way that the $\bar{\underline{x}}_{i,t}$ have mean $\underline{0}_p$. Similarly, we set $\bar{X}_i = X_i - \underline{1}_T\underline{m}'$. Since $\ddot{X}_i = QX_i = Q(\bar{X}_i + \underline{1}_T\underline{m}') = Q\bar{X}_i = \ddot{\bar{X}}_i$, one reduces the analysis to the one made in item viii).

Remark 8. *As a variation of item ii), a simpler argument (not based on the theory of Toeplitz matrices) can be used to prove that all the eigenvalues of the matrix Ψ belong to the interval $\left[1 - \frac{2\rho}{1-\rho}, 1 + \frac{2\rho}{1-\rho}\right]$. This result follows by seeing the matrix Ψ as a perturbation of the identity matrix, then applying Gershgorin's circle theorem Gershgorin, 1931. Indeed, all the eigenvalues of Ψ (which are non-*

[3] We recall that a sequence of random real variables b_T, $T = 1, 2, \ldots$, converges almost surely to $b \in \mathbb{R}$ if $\text{Prob}(\lim_{T\to+\infty} b_T = b) = 1$.

negative since Ψ is symmetric and positive semi-definite) belong to the union of the T Gershgorin's circles $\mathcal{C}_i$ ($i = 1, \ldots, T$) in the complex plane, which have the same center 1 and respective radii $\sum_{j=1,\ldots,T, j\neq i} |\Psi_{ij}|$. The latter radii can be bounded from above by $\frac{2\rho}{1-\rho}$, which follows from a geometric series argument based on Eq. (A.11). We have preferred to use in the main text the argument based on Toeplitz matrices, since imposing $\left[1-\frac{2\rho}{1-\rho}, 1+\frac{2\rho}{1-\rho}\right] \subset (0, +\infty)$ requires the additional assumption $\rho < \frac{1}{3}$ (instead, $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset (0, +\infty)$ holds for any $\rho \in (-1, 1)$). Moreover, such argument produces an even better estimate (when $\sum_{j=1,\ldots,T, j\neq i} |\Psi_{ij}|$ is replaced by its upper bound $\frac{2\rho}{1-\rho}$), since $1-\frac{2\rho}{1-\rho} < \frac{1-\rho^2}{1+\rho^2+2\rho}$ and $1+\frac{2\rho}{1-\rho} = \frac{1-\rho^2}{1+\rho^2-2\rho}$, hence $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset \left[1-\frac{2\rho}{1-\rho}, 1+\frac{2\rho}{1-\rho}\right]$.

Proof of Eq. (2.23)

To make the reading easier, the proof of Eq. (2.23) is divided into several steps.

i) *The following holds:*

$$Q'\Phi^+Q = \Phi^+ . \tag{A.31}$$

This is obtained as follows. First, since the matrix $Q = Q'$ is idempotent, one gets

$$Q'\Phi^+Q = \Phi^+Q \tag{A.32}$$

by Maciejewski and Klein, 1985, Appendix. Additionally, since Moore-Penrose pseudoinversion commutes with transposition (Barata and Hussein, 2012), we get

$$\left(\Phi^+Q\right)' = Q'\left(\Phi^+\right)' = Q'\left(\Phi'\right)^+ = Q'\Phi^+ = \Phi^+ , \tag{A.33}$$

where the last step follows again by Maciejewski and Klein, 1985, Appendix and by the symmetry of Φ. Transposing Eq. (A.33) and combining it with the symmetry of Φ^+ and with Eq. (A.32), we get Eq. (A.31).

ii) *The following decomposition holds*:

$$\ddot{X}_n'\Phi^+\ddot{X}_n = \; X_n'Q'\Phi^+QX_n = X_n'\Phi^+X_n = X_n'\left[\Phi^+ - Q\Psi^{-1}Q'\right]X_n + X_n'Q\Psi^{-1}Q'X_n \quad . \tag{A.34}$$

This is obtained straightforwardly, by applying item i) to get the second equality.

iii) *Under the assumption* $\lim_{T\to\infty} \|\Phi^+ - Q\Psi^{-1}Q'\|_2 = 0$ *stated in Eq. (2.22), the following probability limit holds:*

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} X_n' \left[\Phi^+ - Q\Psi^{-1}Q'\right] X_n = 0_{p\times p}\,. \tag{A.35}$$

This is obtained as follows. Denoting by $\epsilon > 0$ an upper bound on the spectral norm of the matrix $\Phi^+ - Q\Psi^{-1}Q'$ and by $\underline{c}_{n,h}$ the h-th column of X_n, the absolute value of the element in position (h,k) of the matrix $X_n' \left[\Phi^+ - Q\Psi^{-1}Q'\right] X_n$ can be bounded from above as follows:

$$|(X_n' \left[\Phi^+ - Q\Psi^{-1}Q'\right] X_n)_{h,k}| \leq \epsilon \|\underline{c}_{n,h}\|_2 \|\underline{c}_{n,k}\|_2 \leq \frac{1}{2}\epsilon(\|\underline{c}_{n,h}\|^2 + \|\underline{c}_{n,k}\|_2^2)\,, \tag{A.36}$$

where Cauchy-Schwarz inequality has been applied, together with the elementary inequality $|a||b| \leq \frac{a^2+b^2}{2}$, for $a, b \in \mathbb{R}$. Since by the assumption $\lim_{T\to\infty} \|\Phi^+ - Q\Psi^{-1}Q'\|_2 = 0$ stated in Eq. (2.22) one can make ϵ tend to 0 as T tends to $+\infty$, and both $\|\underline{c}_{n,h}\|_2^2$ and $\|\underline{c}_{n,k}\|_2^2$ are summations of T independent and identically distributed random variables with finite mean and finite second order moments, by applying Chebyschev's weak law of large numbers, one gets

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} |(X_n' \left[\Phi^+ - Q\Psi^{-1}Q'\right] X_n)_{h,k}| = 0\,. \tag{A.37}$$

Finally, one gets Eq. (A.35) from Eq. (A.37), since for a sequence of random matrices with fixed dimension, element-wise convergence in probability implies convergence in probability of the whole sequence Herriges, 2011.

iv) *The following probability limit holds:*

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} X_n' Q\Psi^{-1}Q' X_n = \frac{1+\rho^2}{1-\rho^2} \mathbb{E}\left\{ \left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right) \left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)' \right\}. \tag{A.38}$$

This is obtained as follows. Exploiting the symmetry of Q and the following Cholesky factorization (see, e.g. Ruud et al., 2000, Section 19.2)

$$\Psi^{-1} = \left(C_{\text{Chol}}^{-1}\right)' C_{\text{Chol}}^{-1}\,, \tag{A.39}$$

where

$$C_{\text{Chol}}^{-1} = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & \cdots & 0 \\ -\rho & 1 & 0 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & -\rho & 1 \end{bmatrix}, \tag{A.40}$$

one gets

$$X_n' Q \Psi^{-1} Q' X_n = X_n' Q' \Psi^{-1} Q X_n = \ddot{\underline{x}}_{n,1} \ddot{\underline{x}}_{n,1}' + \sum_{t=2}^{T} \left(\frac{1}{\sqrt{1-\rho^2}} (\ddot{\underline{x}}_{n,t} - \rho \ddot{\underline{x}}_{n,t-1}) \right) \left(\frac{1}{\sqrt{1-\rho^2}} (\ddot{\underline{x}}_{n,t} - \rho \ddot{\underline{x}}_{n,t-1}) \right)' . \tag{A.41}$$

Hence, from Eq. (A.41) one gets

$$\begin{aligned} &\operatorname*{plim}_{T \to +\infty} \frac{1}{T} \sum_{n=1}^{N} X_n' Q \Psi^{-1} Q' X_n \\ &= \operatorname*{plim}_{T \to +\infty} \frac{1}{T} \left[\ddot{\underline{x}}_{n,1} \ddot{\underline{x}}_{n,1}' + \sum_{t=2}^{T} \left(\frac{1}{\sqrt{1-\rho^2}} (\ddot{\underline{x}}_{n,t} - \rho \ddot{\underline{x}}_{n,t-1}) \right) \left(\frac{1}{\sqrt{1-\rho^2}} (\ddot{\underline{x}}_{n,t} - \rho \ddot{\underline{x}}_{n,t-1}) \right)' \right] . \end{aligned} \tag{A.42}$$

Now, we compute the probability limit in the right-hand side of Eq. (A.42) by considering separately the following various terms.

iv.a) The following holds:

$$\operatorname*{plim}_{T \to +\infty} \frac{1}{T} \ddot{\underline{x}}_{n,1} \ddot{\underline{x}}_{n,1}' = 0_{p \times p} . \tag{A.43}$$

This is obtained by applying directly Chebyschev's inequality (Ruud et al., 2000, Section D.2), since each element of the matrix $\ddot{\underline{x}}_{n,1} \ddot{\underline{x}}_{n,1}'$ has finite mean and finite second order moments.

iv.b) Similarly, since the addition of a finite number of terms like the one reported in Eq. (A.43) does not change the probability limit, one gets

$$\begin{aligned} \operatorname*{plim}_{T \to +\infty} \frac{1}{T} \sum_{t=2}^{T} \ddot{\underline{x}}_{n,t} \ddot{\underline{x}}_{n,t}' &= \operatorname*{plim}_{T \to +\infty} \frac{1}{T} \sum_{t=2}^{T} \ddot{\underline{x}}_{n,t-1} \ddot{\underline{x}}_{n,t-1}' \\ &= \operatorname*{plim}_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} \ddot{\underline{x}}_{n,t} \ddot{\underline{x}}_{n,t}' \\ &= \operatorname*{plim}_{T \to +\infty} \frac{1}{T} X_n' Q' Q X_n \\ &= \mathbb{E} \left\{ \left(\underline{x}_{n,1} - \mathbb{E} \left\{ \underline{x}_{n,1} \right\} \right) \left(\underline{x}_{n,1} - \mathbb{E} \left\{ \underline{x}_{n,1} \right\} \right)' \right\} , \end{aligned} \tag{A.44}$$

where the last equality is obtained by exploiting the eigendecomposition of $Q'Q$ (which, combined with the assumptions on the $\underline{x}_{n,t}$, shows that each element in position (h,k) of the matrix $X_n' Q' Q X_n$ is the summation of $T-1$ independent random variables with mean $\left(\mathbb{E} \left\{ \left(\underline{x}_{n,1} - \mathbb{E} \left\{ \underline{x}_{n,1} \right\} \right) \left(\underline{x}_{n,1} - \mathbb{E} \left\{ \underline{x}_{n,1} \right\} \right)' \right\} \right)_{h,k}$ and the same finite variance), then applying Chebyshev's weak law of large

numbers.

iv.c) Moreover,

$$\begin{aligned}
&\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\ddot{\underline{x}}_{n,t}\ddot{\underline{x}}'_{n,t-1}\\
=\;&\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\left(\underline{x}_{n,t}-\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)\left(\underline{x}_{n,t-1}-\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)'\\
=\;&\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\left[\underline{x}_{n,t}\underline{x}'_{n,t-1}-\underline{x}_{n,t}\frac{\sum_{\tau=1}^{T}\underline{x}'_{n,\tau}}{T}-\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\underline{x}'_{n,t-1}+\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}\sum_{\tau=1}^{T}\underline{x}'_{n,\tau}}{T^2}\right]\\
=\;&\mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}+\mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}\\
=\;&0_{p\times p}\,,
\end{aligned}\tag{A.45}$$

where the second-last equality comes from the fact that the probability limit of the product of two factors equals the product of the probability limits of the two factors, when the latter probability limits exist (this is a consequence of the Continuous Mapping Theorem Florescu, 2014, Theorem 7.33), and from the assumptions on the $\underline{x}_{n,t}$.

iv.d) Finally, Eq. (A.38) is obtained by combining items iv.a), iv.b), and iv.c), and taking into account the constant factors in Eq. (A.42).

v) *Final part of the proof of Eq. (2.23).*

To conclude, one gets Eq. (2.23) by combining Eqs. (A.34), (A.35), and (A.38), then summing over N.

Discussion of the validity of the assumption $\lim_{T\to\infty}\|\Phi^{+}-Q\Psi^{-1}Q'\|_2=0$

First, we prove a related result. In the following, for a matrix $M\in\mathbb{R}^{T\times T}$, $\|M\|_{HS}=\sqrt{\frac{1}{T}\sum_{i,j=1}^{T}M_{i,j}^2}$ denotes its Hilbert-Schmidt norm Gray, 2006, Eq. (2.17), which is a scaled version of its Frobenius norm $\|M\|_F=\sqrt{\sum_{i,j=1}^{T}M_{i,j}^2}$.

The following holds:

$$\lim_{T\to\infty}\|\Phi^{+}-Q\Psi^{-1}Q'\|_{HS}=0\,.\tag{A.46}$$

Eq. (A.46) is derived by combining several steps, which are listed next, together with pointers to some theoretical results available in the literature that are directly

applied for their proofs, and checks of the assumptions of such results in the context of our analysis. In the following, for a better clarity of exposition of this part, the dependence of Ψ and other matrices on T is highlighted by including the subscript T in the notation.

i) $\lim_{T\to\infty} \|\Psi_T - C_T\|_{HS} = 0$, where C_T is a suitable symmetric and positive definite circulant matrix[4] approximation of the symmetric Toeplitz matrix Ψ_T (application of Gray, 2006, Lemma 4.6 to the circulant matrix approximation C_T of Ψ_T coming from Gray, 2006, Eq. (4.32), where C_T is also symmetric and positive definite due to the symmetry and positive definiteness of the Toeplitz matrix Ψ_T; the application itself of Gray, 2006, Lemma 4.6 is made possible in this case by the convergence of $1 + 2\sum_{k=1}^{+\infty} |\rho|^k$).

ii) $\lim_{T\to\infty} \|\Phi_T - Q_T C_T Q_T'\|_{HS} = 0$ (definition of Φ_T as $\Phi_T = Q_T \Psi_T Q_T'$ and combination of item i) with Gray, 2006, Lemma 2.3 and the fact that $\|Q_T\|_2 = \max_{t=1,\dots,T} |\lambda_t(Q_T)| = 1$).

iii) $\lim_{T\to\infty} \|\Phi_T^+ - (Q_T C_T Q_T')^+\|_{HS} = 0$ (combination of item ii) with Wedin, 1973, Theorem 4.1, made possible by the fact that Φ_T and $Q_T C_T Q_T'$ have the same rank $T-1$, and the spectral norm of Φ_T^+ and the one of $(Q_T C_T Q_T')^+$ are uniformly bounded with respect to T, due respectively to item iv) in the proof of Eq. (2.21) and to the characterization of the eigenvalues of C_T provided in Gray, 2006, Eq. (4.34), combined with $m_f = \frac{1-\rho^2}{1+\rho^2+2\rho} > 0$, see Eq. (A.19)).

iv) $\lim_{T\to\infty} \|\Psi_T^{-1} - C_T^{-1}\|_{HS} = 0$ (application of Gray, 2006, Theorem 5.2 (b) to the function $f(\lambda)$ reported in Eq. (A.19)).

v) $\lim_{T\to\infty} \|Q_T \Psi_T^{-1} Q_T' - Q_T C_T^{-1} Q_T'\|_{HS} = 0$ (obtained likewise item ii)).

vi) $Q_T C_T^{-1} Q_T' = (Q_T C_T Q_T')^+$ (obtained by exploiting the following facts: since C_T is a symmetric matrix, it has the factorization $C_T = U_T \Sigma_T U_T'$ for an orthogonal matrix $U_T \in \mathbb{R}^{T\times T}$ and a diagonal matrix $\Sigma_T \in \mathbb{R}^{T\times T}$ containing its eigenvalues, which are positive; since C_T is also a circular matrix, one can choose one column of U_T to be proportional to $\underline{1}_T$, since this is an eigenvector of C_T Gray, 2006,

[4]We recall that a symmetric circulant matrix is a symmetric Toeplitz matrix (see footnote 1) with $m_t = m_{T-t}$, for $t = 1, \dots, T-1$ Gray, 2006.

Theorem 3.1; Q_T represents the orthogonal projection of $\mathbb{R}^T$ onto its subspace L orthogonal to $\underline{1}_T$; as a consequence of the facts above, one can easily check that $Q_T C_T^{-1} Q_T'$ satisfies all the defining properties[5] Barata and Hussein, 2012 of the Moore-Penrose pseudoinverse of $Q_T C_T Q_T'$, hence $Q_T C_T^{-1} Q_T' = (Q_T C_T Q_T')^+$ by the uniqueness of the Moore-Penrose pseudoinverse Barata and Hussein, 2012).

vii) Finally, Eq. (A.46) is obtained by combining items iii), v), and vi).

Remark 9. *One can easily check (e.g., by a numerical study for selected values of T and of the parameter ρ in Ψ) that, in general, the stronger result $\Phi^+ = Q\Psi^{-1}Q'$ does not hold. This depends on the fact that, given two matrices $M_1, M_2 \in \mathbb{R}^{T\times T}$, typically $(M_1 M_2)^+ \neq M_2^+ M_1^+$, apart from particular cases Dattorro, 2010, Eq. (E.0.0.0.1).*

Eq. (2.22), i.e., $\lim_{T\to\infty} \|\Phi^+ - Q\Psi^{-1}Q'\|_2 = 0$, represents a stronger convergence requirement on $\Phi^+ - Q\Psi^{-1}Q'$ with respect to the convergence result provided by Eq. (A.46), which has been proved above. This depends on the fact that, for any matrix $M \in \mathbb{R}^{T\times T}$, one has $\|M\|_2 \geq \|M\|_{HS}$ Gray, 2006, Eq. (2.19). The validity of Eq. (2.22) has been assumed to complete the proof of Eq. (2.23) - specifically, of part of item iii) therein - since a similar argument based on Eq. (A.46) would be not enough to complete that proof. Although we are currently unable to provide a formal proof of Eq. (2.22) - which is why it has been reported here as an assumption - its validity is strongly supported by the numerical results shown in Figure 26, where the spectral norm error $\|\Phi^+ - Q\Psi^{-1}Q'\|_2$ is reported for several choices of T and ρ (similar results are obtained for a wider range of values of T and other values of ρ). The difficulty in getting a proof of Eq. (2.22) depends on the fact that the vector $\underline{1}_T$ is not an eigenvector of Ψ (although it is an eigenvector of its circulant matrix approximation). Hence, there is no guarantee a priori that all the elements of an orthonormal basis of eigenvectors of Ψ have nonzero orthogonal projections onto $\underline{1}_T$ (indeed, it can be easily checked numerically - e.g., by finding a basis of eigenvectors of Ψ for a few choices of T, then computing such orthogonal projections - that they are typically nonzero)[6].

[5]For example, $(Q_T C_T Q_T')(Q_T C_T^{-1} Q_T')(Q_T C_T Q_T') = (Q_T C_T Q_T')$ is checked by using a common basis of eigenvectors $\underline{e}_i \in \mathbb{R}^T$ (for $i = 1,\ldots,T$) of Q_T and C_T, i.e., by showing that $(Q_T C_T Q_T')(Q_T C_T^{-1} Q_T')(Q_T C_T Q_T')\underline{e}_i = (Q_T C_T Q_T')\underline{e}_i$ for each such eigenvector.

[6]Otherwise, if $\underline{1}_T$ were an eigenvector of Ψ, one could proceed likewise in item vi) of the proof of Eq. (A.46).

This suggests, as a possible way to proceed in the proof, to investigate theoretically if such orthogonal projections converge uniformly to 0 as T tends to $+\infty$. In any case, in this appendix we have reported Eq. (A.46) together with its proof, because such equation is obviously related to Eq. (2.22), and because a proof of the latter could be obtained by combining Eq. (A.46) with specific properties of the matrix Ψ. It is also worth mentioning that such proof would be not necessarily based on the use of a circulant matrix approximation of Ψ, then of Ψ^{-1} (for which negative results on the spectral norm approximation error are known, unless a related but restricted notion of finite-term strong convergence is considered Sun, Jiang, and Baras, 2003, Theorem 1).

Appendix 3: other large-sample approximations of the conditional generalization error, and associated optimization problems

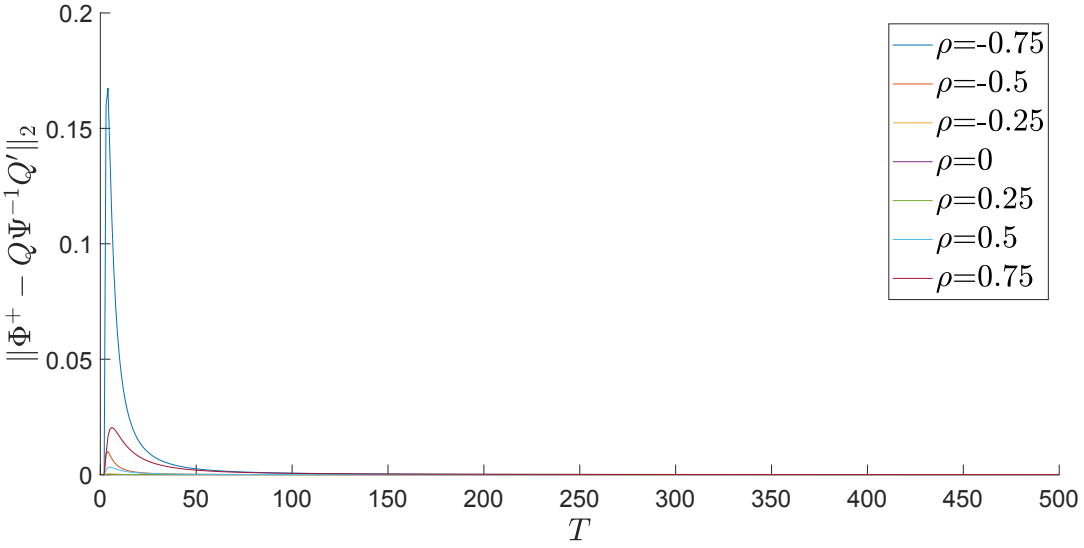
In the following, we report some notes about how the analysis made in Sections 2 and 2 can be modified if one considers, respectively, the case of large N (whose application is potentially of interest in microeconometrics), and the one in which both N and T are large. For simplicity, we limit this extension of the analysis to the case $\rho = 0$, for which one obtains the simplified expressions $\Psi = I_T$ and $\Phi = Q\Psi Q' = QQ' = I_T - \frac{1}{T}\underline{1}_T\underline{1}_T'$. Then, one gets $Q'\Phi^+Q = \Phi^+ = QQ' = Q'Q$ by combining Eq. (A.31) and the relation between the singular value decomposition of a matrix and the singular value decomposition of its Moore-Penrose pseudoinverse.

First, we consider the case in which N is large. Assuming stationarity and mutual independence of different observations associated with the same unit, computations of the elements of the matrix

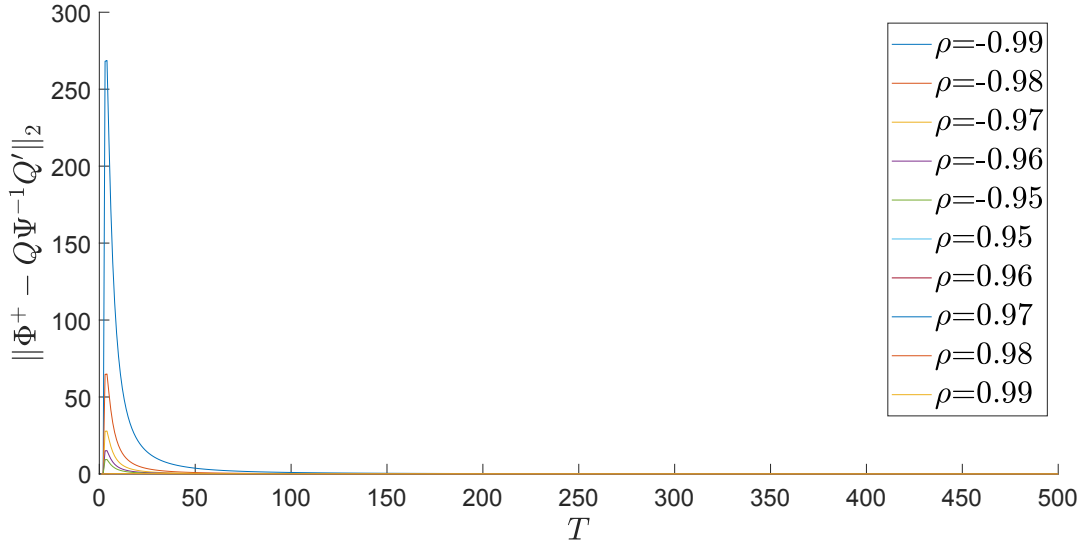
$$\ddot{X}_n'\Phi^+\ddot{X}_n = \ddot{X}_n'Q'\Phi^+Q\ddot{X}_n = \ddot{X}_n'Q'Q\ddot{X}_n = \ddot{X}_n'\ddot{X}_n \tag{A.47}$$

show that

$$\begin{aligned} & \mathbb{E}\left\{\ddot{X}_n'\ddot{X}_n\right\} \\ = \; & \mathbb{E}\left\{\sum_{t=1}^{T}\left(\underline{x}_{n,t} - \frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)\left(\underline{x}_{n,t} - \frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)'\right\} \end{aligned}$$



(a)



(b)

Figure 26: Spectral norm error $\|\Phi^+ - Q\Psi^{-1}Q'\|_2$ as a function of T for (a) several choices of $\rho \in (-1, 1)$, and (b) other choices of $\rho \in (-1, 1)$ near either -1 or 1.

$$
\begin{aligned}
&= \mathbb{E}\left\{\sum_{t=1}^{T}\left(\frac{T-1}{T}\underline{x}_{n,t}-\frac{\sum_{\tau=1\ldots,T,\tau\neq t}\underline{x}_{n,\tau}}{T}\right)\left(\frac{T-1}{T}\underline{x}_{n,t}-\frac{\sum_{\tau=1\ldots,T,\tau\neq t}\underline{x}_{n,\tau}}{T}\right)'\right\}\\
&= T\left(\frac{(T-1)^2}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\underline{x}_{1,t}'\right\}-2\frac{(T-1)^2}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\right\}\mathbb{E}\left\{\underline{x}_{1,t}'\right\}\right.\\
&\quad\left.+\frac{(T-1)}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\underline{x}_{1,t}'\right\}+\frac{(T-1)(T-2)}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\right\}\mathbb{E}\left\{\underline{x}_{1,t}'\right\}\right)\\
&= T\left(\frac{(T-1)T}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\underline{x}_{1,t}'\right\}-\frac{(T-1)T}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\right\}\mathbb{E}\left\{\underline{x}_{1,t}'\right\}\right)\\
&= (T-1)\mathbb{E}\left\{\left(\underline{x}_{n,1}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)\left(\underline{x}_{n,1}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)'\right\}. \qquad \text{(A.48)}
\end{aligned}
$$

Under mild technical conditions (e.g., under the additional assumption that mutual independence extends to all the $\underline{x}_{n,t}$, including those associated with different units, and that all the $\underline{x}_{n,t}$ are identically distributed[7] and have finite moments up to the order 4), from Eq. (A.48) one gets, applying Chebyshev's weak law of large numbers likewise in part of Appendix 2,

$$
\operatorname*{plim}_{N\to+\infty}\frac{1}{N(T-1)}\sum_{n=1}^{N}\ddot{X}_n'\ddot{X}_n = A\,, \qquad \text{(A.49)}
$$

where

$$
A = A' := \mathbb{E}\left\{\left(\underline{x}_{1,1}-\mathbb{E}\left\{\underline{x}_{1,1}\right\}\right)\left(\underline{x}_{1,1}-\mathbb{E}\left\{\underline{x}_{1,1}\right\}\right)'\right\} \qquad \text{(A.50)}
$$

is a symmetric and positive semi-definite matrix. Likewise for what concerns A_N in Section 2, the positive definiteness of A is also assumed in the following.

When (A.49) holds and $\rho = 0$, using also Eqs. (2.18) and (2.19) and the property $Q\underline{1}_T = \underline{0}_T$, one gets the following large-sample approximation with respect to N for the conditional generalization error (2.17), where the dependence on N has been highlighted:

$$
\begin{aligned}
(2.17) \;&\simeq\; \frac{1}{N}\frac{\sigma^2}{(T-1)T^2}\underline{1}_T'X_iA^{-1}X_i'\underline{1}_T+\frac{\sigma^2}{T}\\
&\quad+\frac{1}{N}\frac{\sigma^2}{T-1}\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)'A^{-1}\underline{x}_i^{test}\right\}-\frac{2}{N}\frac{\sigma^2}{(T-1)T}\underline{1}_T'X_iA^{-1}\mathbb{E}\left\{\underline{x}_i^{test}\right\}\\
&= \frac{\sigma^2}{T}+\frac{1}{N}\frac{\sigma^2}{T-1}\mathbb{E}\left\{\left\|A^{-\frac{1}{2}}\left(\frac{1}{T}\left(\underline{1}_T'X_i\right)'-\underline{x}_i^{test}\right)\right\|_2^2\right\}, \qquad \text{(A.51)}
\end{aligned}
$$

[7] This assumption could be relaxed in order to apply in the analysis another suitable form of the weak law of large numbers, valid for the case of dependent/not identically distributed random variables.

where $A^{-\frac{1}{2}}$ is the principal square root of the symmetric and positive definite matrix A^{-1}.

Second, we consider the case in which both N and T are large. In this case, (A.49) is replaced by

$$\operatorname*{plim}_{N,T\to+\infty} \frac{1}{N(T-1)} \sum_{n=1}^{N} \ddot{X}_n' \ddot{X}_n = A\,, \tag{A.52}$$

for the same matrix A as above.

When (2.20) and (A.52) hold and $\rho = 0$, the conditional generalization error (2.17) has the following large-sample approximation with respect to N and T:

$$\begin{aligned} (2.17) \;\simeq\; & \frac{1}{N}\frac{\sigma^2}{T-1}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A^{-1}\mathbb{E}\left\{\underline{x}_{i,1}\right\} + \frac{\sigma^2}{T} \\ & + \frac{1}{N}\frac{\sigma^2}{T-1}\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)' A^{-1}\underline{x}_i^{test}\right\} - \frac{2}{N}\frac{\sigma^2}{T-1}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A^{-1}\mathbb{E}\left\{\underline{x}_i^{test}\right\} \\ = \;& \frac{\sigma^2}{T} + \frac{\sigma^2}{N(T-1)}\mathbb{E}\left\{\left\| A^{-\frac{1}{2}}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\} - \underline{x}_i^{test}\right)\right\|_2^2\right\}. \end{aligned} \tag{A.53}$$

Starting from the large-sample approximations (A.51) and (A.53) for the conditional generalization error, and adopting the model (2.27) for the variance σ^2, two optimization problems similar to (2.28) can be stated and solved. For simplicity, in the following we make some approximations in the analysis of their optimal solutions.

In the first problem, one optimizes the corresponding large-sample approximation of the conditional generalization error with respect to N (or equivalently, with respect to c, as in (2.28)), whereas T is fixed. More precisely, for C sufficiently large (in such a way that the large-sample approximation (A.51) can be assumed to hold for every $c \in [c_{\min}, c_{\max}]$) and under the approximation $NTc \simeq C$ at optimality[8], setting

$$K_i' \;:=\; \mathbb{E}\left\{\left\| A^{-\frac{1}{2}}\left(\frac{1}{T}\left(\underline{1}_T' X_i\right)' - \underline{x}_i^{test}\right)\right\|_2^2\right\}, \tag{A.54}$$

the first optimization problem can be written as

$$\begin{aligned} & \operatorname*{minimize}_{c\in[c_{\min},c_{\max}]} \quad \left(\frac{kc^{-\alpha}}{T} + K_i' \frac{kc^{-\alpha}}{\frac{C}{c}\left(1-\frac{1}{T}\right)}\right) \\ = \; & \operatorname*{minimize}_{c\in[c_{\min},c_{\max}]} \quad \frac{1}{T}\left(kc^{-\alpha} + \frac{T^2 K_i' k}{C(T-1)} c^{1-\alpha}\right), \end{aligned} \tag{A.55}$$

[8]This follows from the fact that the large-sample approximation (A.51) of the conditional generalization error is a decreasing function of N, for each fixed choice of the measurement noise variance σ^2, hence for each choice of c.

whose optimal solution c° has the following expression:

i) if $0 < \alpha < 1$ ("decreasing returns of scale") and

   (a) $c^\star := \frac{C(T-1)\alpha}{K_i' T^2 (1-\alpha)} \in [c_{\min}, c_{\max}]$: $c^\circ = c^\star$;

   (b) $c^\star < c_{\min}$: $c^\circ = c_{\min}$;

   (c) $c^\star > c_{\max}$: $c^\circ = c_{\max}$;

ii) if $\alpha > 1$ ("increasing returns of scale"): $c^\circ = c_{\max}$;

iii) if $\alpha = 1$ ("constant returns of scale"): $c^\circ = c_{\max}$.

The analysis of the second problem (whose optimization variables are c, N, and T, as the large-sample approximation of the conditional generalizarion error with respect to both N and T is optimized) is slightly more involved, since it is formulated in terms of a larger number of optimization variables. Nevertheless, solving such problem can be reduced to solving, for each c, an optimization subproblem in which the same objective function is minimized with respect to the pair (N,T). In this problem, admissible such pairs have to satisfy the constraint $NTc \leq C$, and also two additional lower bounds $N \geq N_{\min} > 0$ and on $T \geq T_{\min} > 0$, under which the large-sample approximation made in (A.53) can be assumed to hold. More precisely, for C sufficiently large and under the approximations $T-1 \simeq T$ and $NTc \simeq C$ at optimality, setting

$$K_i'' := \mathbb{E}\left\{ \left\| A^{-\frac{1}{2}} \left(\mathbb{E}\left\{ \underline{x}_{i,1} \right\} - \underline{x}_i^{test} \right) \right\|_2^2 \right\}, \tag{A.56}$$

the second optimization problem can be written as

$$\begin{aligned} & \underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}} && \left(\frac{k c^{-\alpha}}{T} + K_i'' \frac{k c^{-\alpha}}{\frac{C}{c}} \right) \\ & \text{s.t.} && \frac{C}{Tc} \geq N_{\min}, T \geq T_{\min}, \\ = & \underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}} && \left(\frac{k}{T} c^{-\alpha} + \frac{K_i'' k}{C} c^{1-\alpha} \right) \\ & \text{s.t.} && \frac{C}{Tc} \geq N_{\min}, T \geq T_{\min}, \end{aligned} \tag{A.57}$$

whose optimal solutions c° have the following expressions (the optimal T is $T^\circ \simeq \frac{C}{N_{\min} c^\circ}$):

i) if $0 < \alpha < 1$ ("decreasing returns of scale"): $c^{\circ} = c_{\min}$;

ii) if $\alpha > 1$ ("increasing returns of scale"): $c^{\circ} = c_{\max}$;

iii) if $\alpha = 1$ ("constant returns of scale"): $c^{\circ} =$ any cost c in the interval $[c_{\min}, c_{\max}]$.

3 Appendix for Chapter 3

3.1 Literature review table

The following table summarizes the literature's findings on the relationship between market size and innovation in the pharmaceutical industry. In particular, the focus is on relevant works coming after Acemoglu and Linn (2004). The reason for such a choice is that Acemoglu and Linn (2004) represents a milestone in investigating the relation between market size and innovation in pharmaceuticals. It overcomes issues emerging in previous studies (such as, and above all, the one of endogeneity) and is taken as a reference point by authors willing to further dig into such a literature stream.
Furthermore, the literature review reports the relationship between market size and innovation in Pharmaceutical Industry only. Indeed, different industries have different definitions of recalls.

Table 23: The table reports relevant papers after Acemoglu and Linn, 2004; NME stands for New Molecular Entities; NDA stands for New Drug Approval.

Paper	Data and sample	Unit of observation	Measurement of innovation	estimation method	report estimate of size	proxy market size
Acemoglu and Linn, 2004	US; March CPS, 1965–2000; March CPS, 1965–2000; FDA; OECD	report number of units	NME [b]	QML	> 0 [4]	demographic measures
Cerda, 2007	US; FOIA request; RND process[1]; U.S statistical abstract[2]; 1968-1997	15 drug categories [1]	NME	FE, GLS, IV, Tobit	>0 [1]	demographic measures
Rake, 2017	U.S.; RND; FDA; OECD; ClinicalTrials.gov[3]; 1974-2008	disease	NDA; NME; Phase II and Phase III trials	QMLE (Poisson, 1995)	0.3444 (NME); 0.3521 (NDA)	demographic measures
Dubois et al., 2015	14 countries [5]; 1997-2007;IMS, WHO	chemical entity; dummies for ATC-1 and ATC-2	NCE (elasticity) [b]	OLS,2SLS,CF approach (Wooldr.,2002)	0.23 (average across ATC classes)	deaths and GDP [5]
Blume-Kohout and Sood, 2013	US; 1998-2010; Pharmaprojects [6]; MEPS; OECD; NIH	49 therapeutic classes	R&D [6]	Negative Bin.; Poisson	0.26; 0.41; 0.51 [7]	demographic shifts [5] [c]

[1] RND process of the pharmaceutical sector (gov. funds); > 0 means that the exogenous increase in market size is initially associated with approximately 0.08 more drugs introduced in the market. These new drugs reduce the mortality rates of individuals aged 65 and older by 0.8 percent. This decrease in mortality rate leads to increases in market size (more demand), producing an additional increase of drugs equal to 0.096

[2] (population data for market size)

[3] Both Cerda and Rake consulted the 19th edition of the Drug Information Handbook published by Lexi-Comp and the American Pharmaceutical Association (Lacy et al., 2010). This handbook is comparable to a pharmaceutical dictionary, providing a list of drugs' active ingredients, the medical conditions the drug is used for, and further information such as adverse effects. The work takes into account only those medical conditions which can be found on the FDA-approved label. Hence, unlabeled and investigational uses are not present. For the period 1974 to 2008, FDA approved 599 unique NMEs and 1,665 unique NDAs. These approvals refer to the 208 diseases or medical indications analyzed in this study. However, an NME or NDA may be used as therapy for several medical indications. In this case, an NME or NDA is counted as innovation for all the medical indications for which it is approved

[4] The estimates suggest that a 1 percent increase in the potential market size for a drug category leads to a 6 percent increase in the total number of new drugs entering the U. S. market.

[5] Data come from IMS (Intercontinental Marketing Services) and include all product sales in 14 countries [a] (Australia, Brazil, Canada, China, France, Germany, Italy, Japan, Mexico, Korea, Spain, Turkey, United Kingdom, USA). Dubois et al. have data on the ATC-4 (they report 607 different classes), the main active ingredient of the drug (they report 6216 different active ingredients), the name of the firm producing the drug, whether it has been licensed, the patent start date, and the format of the drug (the work reports 471 different formats). Products in the same ATC-4 by definition have the same indication and mechanism of action. The authors do not consider OTC drugs. Quantities are given in standard units, one standard unit corresponding to the smallest typical dose of a product form, as defined by IMS Health.

[6] Pharmaprojects trend data "snapshot"; (focus on R&D): focus only on one instance of innovation as explained in B. H. Hall and Rosenberg, 2010. Authors specify the adoption of clinical trials (from pre-clinical Phase to Phase III) not taken from ClinicalTrials.gov (see below)

[7] For a drug class with average Medicare market share (41%, in 2004–2005), Duggan and Scott Morton's result translates to an 11% increase in revenues following Medicare Part D. Our Phase I estimates correspond, for a drug class with average Medicare market share, to a 26% increase for 2004–2005, a 33% increase post-implementation in 2006–2007, and a lagged 51% increase in 2008–2010. These estimates imply an elasticity of Phase I clinical trials of 2.4 to 4.7 compared to the market size, bracketing Acemoglu and Linn's estimated elasticity of 3.5 for approved new molecular entities (NMEs). However, when considering all clinical trials combined—including Phase III trials for supplemental indications the estimated elasticity of clinical trials with respect to market size is somewhat lower than Acemoglu and Linn's estimated elasticity of 6 for all new drug approvals, but certainly still more prominent than the Dubois et al. (2011) estimate of about 0.25. Summary results: "The results indicate that the increase in outpatient prescription drug coverage provided through Medicare Part D has had a significant impact on pharmaceutical R&D "

Critiques:

[a] Blume-Kohout and Sood, 2013 states that several of the countries chosen regulate prescription drug prices, and regulations may change rapidly over time. Thus, given the lower expected profit per consumer and greater uncertainty about future profits and prices, firms' R&D decisions are likely to be less responsive to a unit change in expected revenues for all these countries combined versus the same unit change in the U.S. market (Sood et al., 2009).

[b] Blume-Kohout and Sood, 2013: they measured firms' innovative activities via clinical trials, whereas Dubois et al. (2011) and Acemoglu and Linn (2004) evaluate the responsiveness of approved and marketed drugs to changes in market size

[c] Dubois et al., 2015: the authors recognize to Blume-Kohout and Sood, 2013 the fact of having exploited an innovative measure of Market Share (policy change in Medicare Part D)

List of controls:

Acemoglu and Linn, 2004 Potential Supply-Side Determinants of Innovation (changes in scientific incentives); Proxies for pre-existing time trends across sectors; lag dependent var; life-years lost; public funding; pre-existing trends; major category trends; health insurance market size; (see page 1077-1080 for further details on variables)

Cerda, 2007: Gov. expenditure (Medicare and social security); Gov. research efforts (grants on research); year dummies; some demographic information such as prevalence rates of disease i on males (fraction of males/white/married attending hospital due to i), blacks, whites, and married individuals as well as the average age of individuals affected by disease i.

Rake, 2017 The empirical analysis draws upon the literature concerning the "demand-pull" versus "technology-push" debate and takes into account demand- and supply-side factors as the explanatory variables for pharmaceutical innovation. Regressors used comprise knowledge stock (consisting of the scientific publications (Pub_{it}) related to medical indication i and published in year t (BioPharmInsight database); Regulatory stringency (average time between the submission of a new drug approval to the FDA and its final approval); pre-sample mean of new pharmaceuticals; mortality rate per medical indication in 1983 to account for differences in the pre-sample prevalence of medical indication; pre-sample technological opportunities are constructed as the average annual growth rate of the knowledge stock from 1979 to 1983.

Blume-Kohout and Sood, 2013 prescription drugs; funding grants for each disease class

3.2 Time to event analysis

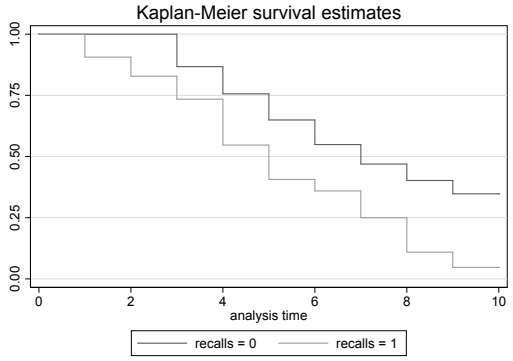


Figure 27: Kaplan-Meyer time to event analysis. The x-axis represents the number of years until death. Recalled products (recalls =1) are already scaled down at 2 years of survival time with respect to not recalled products (recalls =0). The median survival time for recalled medicines is about 4 years, while the one for not-recalled medicines is about 7 years. The survival function of recalled products persists in its falling below the survival function of not recalled drugs. This means that recalls affect sales for a long period of time. In other words, within the market of the recalled product there will be a lack of potential sales left by the recalled products. Missed sales are hence not a temporary event, demonstrating the length of the lack that should be covered to fill the gap provoked by the product's recall.

3.3 Abnormal Values (firm and product levels)

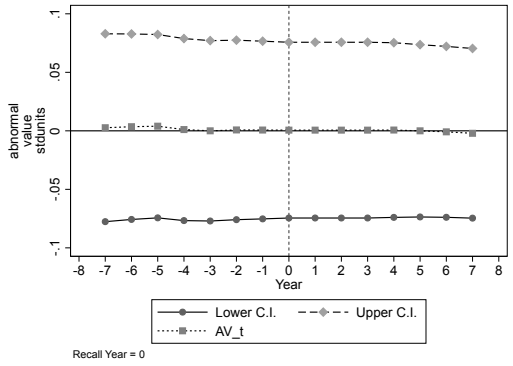
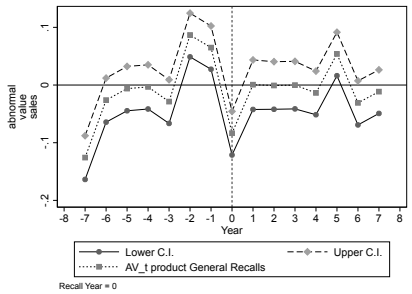
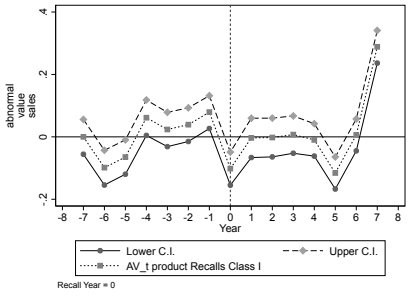


Figure 28: Effect of recalls on market sales once firms having undergone a major recall are cancelled out. The absence of any effect (i.e. increases of sales due to recalls of competitors) at recall time for products other than the ones of the recalled firm witnesses the absence of compensations both at time 0 or soon after the recall.

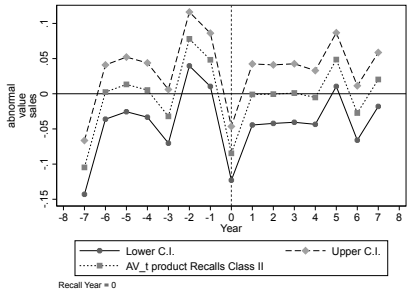
The following figures represent abnormal values at firm and product level for different typologies of recall (according to their gravity).



Product: general recalls sales

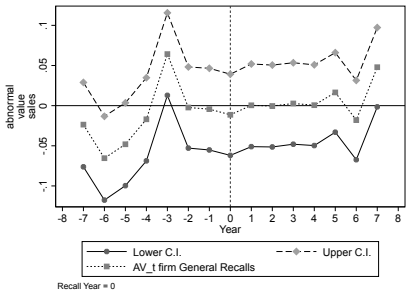


Product: Class I recalls sales

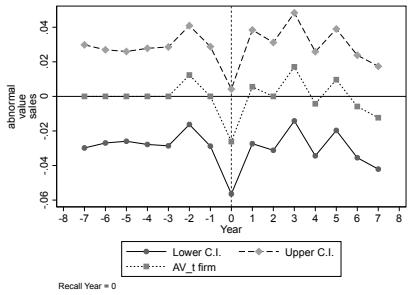


Product: Class II recalls sales

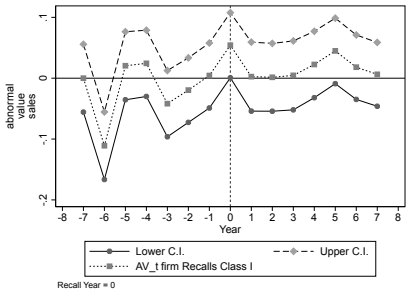
Figure 29: Abnormal values for product aggregation. Years are normalized. Year 0 represents the year of recall. The three scenarios include the path of sales before and after the recall year, using three different definitions of recalls: Class I recalls, general recalls and Class II recalls. As shown in the pictures, sales at product level drop at recall year.



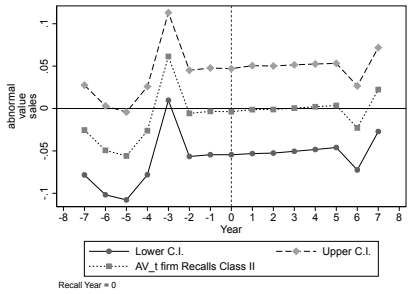
Firm: general recalls sales



Firm: Maj.recalls sales



Firm: Class I recalls sales



Firm: Class II recalls sales

Figure 30: Abnormal values at firm aggregation. Years are normalized. Year 0 represents the year of recall. The four scenarios include the path of sales before and after the recall year, using four different definitions of recalls: major recalls, Class I recalls, general recalls and Class II recalls. A part from major recalls, catching firms unaware, the other types of recalls do not affect firms sales. This might be due to compensation of sales within firms.

The effect of recalls is evident for all aggregations but firm level, where the effect is not evident (future development).Possible hypotheses are detailed in the main text.

3.4 Firm and product levels

Summary statistics Literature focused in the first instance on the impact of firm size on innovation. In 1942, Schumpeter indicated that larger firms are more innovative than smaller firms, and since then, firm size has become one of the most often investigated innovation determinants (Kolluru and Mukhopadhaya, 2017). The debate mainly concerned the role of downsizing in firms and specifically whether downsizing could stimulate innovation or not. On the one hand, some innovation economists believe that the bigger the firms, the more the resources at their disposal,and the higher the probability of innovating. For such a stream of literature, downsizing with layoffs and reduction of organizational slacks may leave the firm at a wrong size, thus, negatively affecting firm performance (Fisher and White, 2000; Hashi and Stojčić,

2013; Luan, Tien, and Chi, 2013 and others.).

On the other hand, some authors support the hypothesis that smaller firms, above all entrants, are more encouraged to innovate to compete with bigger established firms. For this stream, elimination of positions and management layers by downsizing may create an internal environment favorable to the generation and survival of new innovative ideas (Ross, 1974; Baumol, Blinder, and Wolff, 2003). However, existing empirical approaches may be flawed by the endogeneity of size: size, indeed, can be related to a firm's unobservable characteristics that in turn affect innovation (Symeonidis, 1996). Firms could yet influence their size and innovation by being active decision-makers, prone to unobservable variables such as propensity to innovate or risk. Authors aware of the endogeneity issue tried to employ diverse strategies to correct it (see among others: Alsharkas, 2014; Stock, Greis, and Fischer, 2002) but just partially succeeded. Indeed controversial results emerged. At market level (Acemoglu and Linn, 2004; Pammolli, Magazzini, and Riccaboni, 2011; Stoneman, 2010; Dubois et al., 2015 and others) the endogeneity issue, as above described, seems to be naturally mitigated with respect to firm-level by the exact definition of the market as a collection of products.

Below are reported the summary statistics at firm and product level. From the results emerges how recalls are more common in old established firms where, however, the effect is null due to compensations and endogeneity (F., 2021).

Table 24: Summary statistics at product and firm level. Databases at product and firm level are unbalanced.

Variable		Prod.			Firm		
		Full Sample	Subs. recalls	Subs. no recalls	Full Sample	Subs. recalls	Subs. no recalls
Sales (log)	Overall mean	12.868	15.872	12.844	14.509	19.808	14.146
	Overall Std. Dev.	3.791	3.674	3.787	4.064	3.278	3.826
	Between Std. Dev	3.657	3.367	3.653	4.080	3.565	3.873
	Within Std. Dev	1.764	1.957	1.764	1.459	1.080	1.486
Age prod./ firm	Overall mean	11.910	13.588	11.909	17.880	31.022	16.850
	Overall Std. Dev.	10.692	11.718	10.687	14.849	15.805	14.024
	Between Std. Dev	9.977	10.821	9.974	13.901	16.723	13.110
	Within Std. Dev	2.748	3.042	2.746	3.022	3.385	2.998
New mol. market (%)	Overall mean	.085	.079	.0851	.	.	.
	Overall Std. Dev.	.279	.271	.279	.	.	.
	Between Std. Dev	.213	.109	.213	.	.	.
	Within Std. Dev	.256	.262	.256	.	.	.
Avg. age prod. by firm/ATC	Overall mean	.	.	.	11.002	11.581	10.921
	Overall Std. Dev.	.	.	.	8.974	6.039	9.130
	Between Std. Dev	.	.	.	8.602	5.858	8.685
	Within Std. Dev	.	.	.	2.985	2.656	3.018
Share generics by firm/ATC	Overall mean	.	.	.	.665	1.534	.623
	Overall Std. Dev.	.	.	.	3.366	3.963	3.361
	Between Std. Dev	.	.	.	3.407	4.486	3.386
	Within Std. Dev	.	.	.	1.668	1.637	1.685
Outflow rate	Overall mean	.	.	.	.086	.085	.086
	Overall Std. Dev.	.	.	.	.226	.148	.231
	Between Std. Dev	.	.	.	.234	.108	.238
	Within Std. Dev	.	.	.	.186	.124	.190

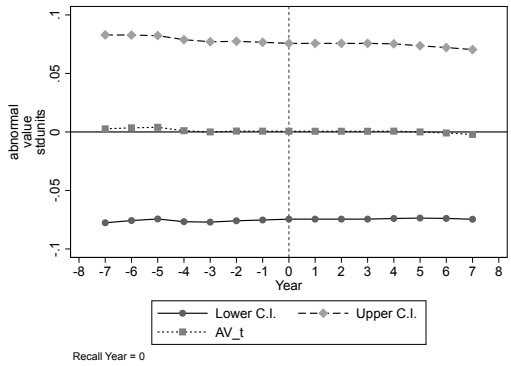
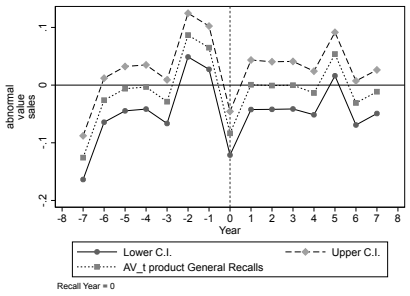
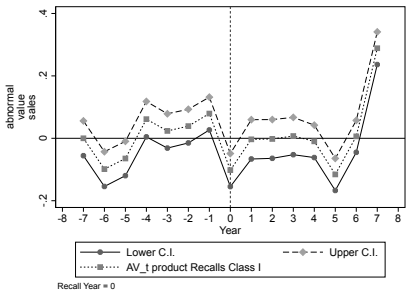


Figure 31: Effect of recalls on market sales once firms having undergone a major recall are cancelled out. The absence of any effect (i.e. increases of sales due to recalls of competitors) at recall time for products other than the ones of the recalled firm witnesses the absence of compensations both at time 0 or soon after the recall.

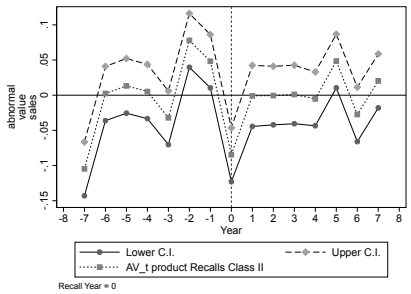
Abnormal Values (firm and product levels) The following figures represent abnormal values at firm and product level for different typologies of recall (according to their gravity).



Product: general recalls sales

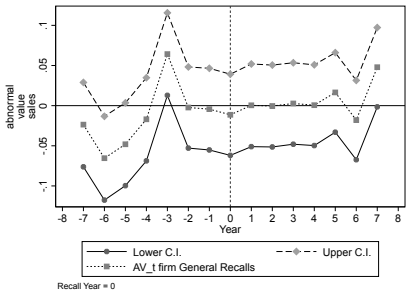


Product: Class I recalls sales

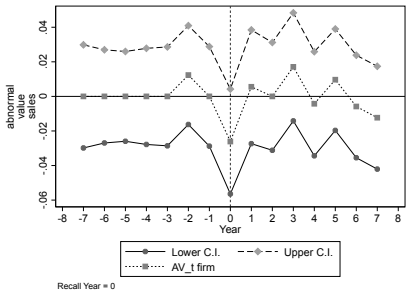


Product: Class II recalls sales

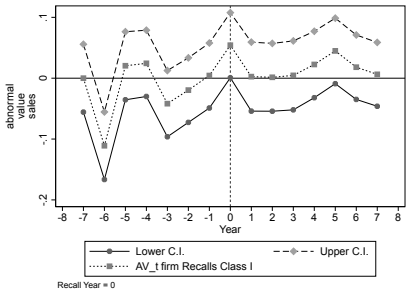
Figure 32: Abnormal values for product aggregation. Years are normalized. Year 0 represents the year of recall. The three scenarios include the path of sales before and after the recall year, using three different definitions of recalls: Class I recalls, general recalls and Class II recalls. As shown in the pictures, sales at product level drop at recall year.



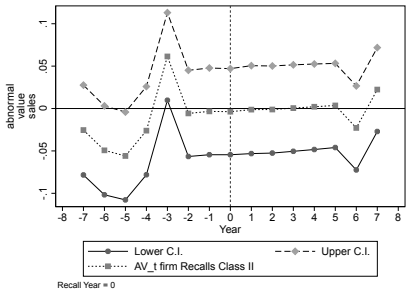
Firm: general recalls sales



Firm: Maj.recalls sales



Firm: Class I recalls sales



Firm: Class II recalls sales

Figure 33: Abnormal values at firm aggregation. Years are normalized. Year 0 represents the year of recall. The four scenarios include the path of sales before and after the recall year, using four different definitions of recalls: major recalls, Class I recalls, general recalls and Class II recalls. A part from major recalls, catching firms unaware, the other types of recalls do not affect firms sales. This might be due to compensation of sales within firms.

The effect of recalls is evident for all aggregations but firm level, where the effect is not evident (future development).Possible hypotheses are detailed in the main text.

Table 25: First stage at different levels

	(Prod. Aggregation)	**(Firm Aggregation)**	**(ATC-Firm Aggregation)**
	Log sales	Log sales	Log sales
recalls	−0.0680		
	(0.2544)		
recalls_{t-1}	−1.6269***		
	(0.2783)		
$\tilde{recalls}$		0.0042	−0.0053
		(0.0033)	(0.0033)
$\tilde{recalls}_{t-1}$		−0.0008	−0.0267**
		(0.0046)	(0.0083)
Age firm		−0.0098	
		(0.0123)	
$\frac{K_{t+1}}{P_{-1}}$		−0.5957***	
		(0.1272)	
prods.		0.0214**	
		(0.0077)	
share generics in ATC		−0.0304	
		(0.0235)	
Year Dummies	Yes	Yes	Yes
Obs.	94567	5787	48915
Groups	17947	892	8634

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust standard errors are in parentheses. (1) fits a F.E. model with at product level. (2) fits a F.E. model at firm level. (4) fits a F.E. model at ATC-Firm level, i.e. ATC lines of productions within firms. $\tilde{recalls}$ represent recalls normalized. Recalls normalized are not employed in the first column since there is no aggregation at product level of analysis. At firm and ATC level recalls are respectively normalized for the number of products within a firm and within an ATC.

First stage rob. checks In this section are displayed the significant coefficients of the first stage employing the number of patients as measure for market size.

Table 26: First stage of the robustness check using the number of patients as measure of market size

	(1)
	# patients
$\tilde{recalls}$ 0.0519	
	(0.0333)
$\tilde{recalls}_{t-1}$	−0.262*
	(0.149)
Year Dummies	Yes
Obs.	1056
Groups	132

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust standard errors are in parentheses. (1) first stage results when MEPS database is employed and market size is measured with the number of patients. Second stage results are in Tab. 14.

4 Appendix for Chapter 4

Matrix completion via nuclear norm regularization of the reconstruction error Given a subset of observed entries of a matrix $\mathbf{A} \in \mathbb{R}^{C \times P}$, Matrix Completion (MC) works by finding a suitable low-rank approximation (say, with rank R) of $\mathbf{A}$, by assuming the following model:

$$\mathbf{A} = \mathbf{C}\mathbf{G}^\top + \mathbf{W}, \tag{A.58}$$

where $\mathbf{C} \in \mathbb{R}^{C \times R}$, $\mathbf{G} \in \mathbb{R}^{P \times R}$, whereas $\mathbf{W} \in \mathbb{R}^{C \times P}$ is a matrix of modeling errors. The rank-R approximating matrix $\mathbf{C}\mathbf{G}^\top$ is found by solving a suitable optimization problem (provided by Eq. (4.2), in the case of the present article). Eq. (A.58) can be written element-wise as $A_{c,p} = \sum_{r=1}^{R} C_{c,r} G_{p,r} + W_{c,p}$. A common interpretation of this equation is as follows (see, e.g., the application of MC to collaborative filtering for movie ratings Trevor Hastie, R. Tibshirani, and Wainwright, 2019). The number $C_{c,r}$ can be interpreted as the degree of membership of row c of matrix $\mathbf{A}$ to some "latent" cluster r (for a total of R such clusters), and $G_{p,r}$ as the prediction of an element in column p of matrix $\mathbf{A}$, conditioned on its row c belonging to cluster r. It is worth mentioning that such an interpretation holds regardless of the signs of the elements $C_{c,r}$ and $G_{p,r}$. As an example, in the case of collaborative filtering for movie ratings, c denotes a specific person, p a specific movie, whereas r may be interpreted as a specific movie genre.

In this work, MC is formulated via the optimization problem (4.2). The objective function of this optimization problem is the sum of two terms: the first one refers to the reconstruction error of the known portion of the matrix, whereas the second one is a regularization term, which biases the reconstructed matrix to have a small nuclear norm. The regularization constant λ controls the trade-off between fitting the known entries of the matrix $\mathbf{A}$ and achieving a small nuclear norm. The latter requirement is often related to getting a low rank of the optimal matrix $\mathbf{Z}^\circ$, which follows by geometric arguments similar to the ones typically adopted to justify how the classical LASSO (Least Absolute Shrinkage and Selection Operator) penalty term achieves effective feature selection in linear regression R. Tibshirani, 1996.

The MC optimization problem (4.2) can be also written as

$$\underset{\mathbf{Z}\in\mathbb{R}^{C\times P}}{\text{minimize}} \left(\frac{1}{2} \|\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{A}) - \mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{Z})\|_F^2 + \lambda \|\mathbf{Z}\|_* \right), \tag{A.59}$$

where, for a matrix $\mathbf{Y} \in \mathbb{R}^{C\times P}$, $(P_{\Omega^{\text{tr}}}(\mathbf{Y}))_{c,p} := Y_{c,p}$ if $(c,p) \in \Omega^{\text{tr}}$, otherwise it is equal to 0. Here, $P_{\Omega^{\text{tr}}}(\mathbf{Y})$ represents the projection of $\mathbf{Y}$ onto the set of positions of observed entries of the matrix $\mathbf{A}$, and $\|\mathbf{Y}\|_F$ denotes the Frobenius norm of $\mathbf{Y}$ (i.e., the square root of the summation of squares of all its entries).

The MC optimization problem (A.59) can be solved by applying the following Algorithm 1, named Soft Impute Mazumder, Trevor Hastie, and R. Tibshirani, 2010b (compared to the original version, here we have included a maximal number of iterations N^{it}, which can be helpful to reduce the computational effort when one has to run the algorithm multiple times, e.g., for several choices of the training set Ω^{tr} and of the regularization constant λ, as in the present work):

Algorithm 1: Soft Impute Mazumder, Trevor Hastie, and R. Tibshirani, 2010b

---

Input: Partially observed matrix $\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{A})$, regularization constant $\lambda \geq 0$, tolerance $\varepsilon \geq 0$, maximal number of iterations N^{it}

Output: Completed matrix $\mathbf{Z}_\lambda \in \mathbb{R}^{C\times P}$

i) Initialize $\mathbf{Z}$ as $\mathbf{Z}^{\text{old}} = \mathbf{0} \in \mathbb{R}^{C\times P}$

ii) Repeat for at most N^{it} iterations:

(a) Set $\mathbf{Z}^{\text{new}} \leftarrow \mathbf{S}_\lambda \left(\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{A}) + \mathbf{P}^{\perp}_{\Omega^{\text{tr}}}(\mathbf{Z}^{\text{old}}) \right)$

(b) If $\frac{\|\mathbf{Z}^{\text{new}} - \mathbf{Z}^{\text{old}}\|_F^2}{\|\mathbf{Z}^{\text{old}}\|_F^2} < \varepsilon$, exit

(c) Set $\mathbf{Z}^{\text{old}} \leftarrow \mathbf{Z}^{\text{new}}$

iii) Set $\mathbf{Z}_\lambda \leftarrow \mathbf{Z}^{\text{new}}$

In Algorithm 1, for a matrix $\mathbf{Y} \in \mathbb{R}^{C\times P}$, $\mathbf{P}^{\perp}_{\Omega^{\text{tr}}}(\mathbf{Y})$ represents the projection of $\mathbf{Y}$ onto the complement of Ω^{tr}, whereas $\mathbf{S}_\lambda(\mathbf{Y}) := \mathbf{U}\mathbf{\Sigma}_\lambda\mathbf{V}^\top$, being $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (with $\mathbf{\Sigma} = \text{diag}[\sigma_1, \ldots, \sigma_R]$) the singular value decomposition of $\mathbf{Y}$, and $\mathbf{\Sigma}_\lambda := \text{diag}[(\sigma_1 - \lambda)_+, \ldots, (\sigma_R - \lambda)_+]$, with $t_+ := \max(t, 0)$.

It is worth mentioning that a particularly efficient implementation of the operator $\mathbf{S}_\lambda(\cdot)$ is possible (by means of the MATLAB function `svt.m`Li and Zhou, 2017), which

is based on the determination of only the singular values σ_i of $\mathbf{Y}$ that are higher than λ, and of their corresponding left-singular vectors $\mathbf{u}_i$ and right-singular vectors $\mathbf{v}_i$. Indeed, all the other singular values of $\mathbf{Y}$ are annihilated in $\mathbf{\Sigma}_\lambda$.

A final remark has to be made about the trade-off between prediction capability and biasedness of MC. Biasedness in MC depends, among others issues, on the way the selection of unobserved entries is made Foucart et al., 2017; Ma and G. H. Chen, 2019 (in the specific case of our application of MC to the discretized **RCA** matrix, only entries belonging to a suitable subset of rows of the matrix $\mathbf{A}$ are obscured). For some MC algorithms, de-biasing is possible Foucart et al., 2017, and can even improve prediction capability. Nevertheless, in general biasedness can be beneficial to prediction capability, due to the well-known trade-off between bias and variance Tibshirani Hastie and R. Tibshirani, 2009. In the particular case of MC achieved via the Soft Impute algorithm, biasedness can be ascribed also to the presence of the regularization constant λ (indeed, for both $\lambda \to 0$ and $\lambda \to +\infty$, the predictions of the optimal solution to the optimization problem (A.59) tend to 0 for the unobserved entries), and to the fact that the Soft Impute algorithm is initialized by a matrix with all entries equal to 0, and terminated at most after a given number of iterations.

Technical details on the construction of the matrix $\mathbf{A}$ and on the the application of the Soft Impute algorithm This subsection details the construction of the matrix $\mathbf{A}$ for our specific problem. As a first step, we removed from the **RCA** matrix its rows associated with countries having less than 5 million inhabitants. Then, the remaining entries of the **RCA** matrix were encoded into 9 groups according to increasing percentiles in the distribution of RCA values. This pre-processing step was done in order to make the elements of the resulting matrix $\mathbf{A} \in \mathbb{R}^{119\times1243}$ of the same order of magnitude. In particular, we defined 4 negative groups ("-4", "-3", "-2", "-1"), representing the case $0 \leq RCA < 1$ (with the group "-4" being the one associated with the lowest values in the RCA distribution) and 4 positive groups ("1", "2", "3", "4"), representing the case $RCA \geq 1$ (with the group "4" being the one associated with the highest values in the RCA distribution)[9]. Originally NaN RCA values were included

[9]As already reported in the Appendix, MC is biased towards low absolute values. To take this into account, we constructed groups that were symmetrically distributed around zero, as the final goal was to discriminate between RCA values respectively lower than 1, and larger than or equal to 1.

in the remaining group "0". In our application of MC, the elements in this group "0" were included neither in the training set, nor in the validation/test set, since no ground truth was available for them.
For computational efficiency reasons, we combined the original MATLAB implementation of Soft Impute (Mazumder, Trevor Hastie, and R. Tibshirani, 2010b) with the MATLAB function `svt.m` (Li and Zhou, 2017). The tolerance of the algorithm was chosen as $\varepsilon = 10^{-9}$. Its number of iterations was set to $N^{\text{it}} = 1500$. The regularization parameter λ was sampled 30 times uniformly on the closed interval $[-1, 15]$ in a logaritmic scale with base 2. A post-processing step was included in MC, thresholding to -4 any element (when present) whose MC reconstruction was lower than -4, and to 4 any element (when present) whose MC reconstruction was higher than 4.

Generalized economic complexity index and related economics complexity indices The GENeralised Economic comPlexitY (GENEPY) index is a recently-introduced economic complexity index (Sciarra, Chiarotti, Ridolfi, et al., 2020b), which can be applied to assess the complexity of both countries and products. It is based on a multidimensional representation of their complexity, which makes it possible to combine, in a single index, the different features of some previously-developed one-dimensional economic complexity indices: the Fitness (F) for countries and Quality (Q) for products, both computed by the Fitness and Complexity (FC) algorithm (Tacchella et al., 2012), and the Economic Complexity Index (ECI) for countries and Product Complexity Index (PCI) for products, both obtained by the earlier Method of Reflections (MR) (Hidalgo and Hausmann, 2009). Each of the latter methods is typically able, indeed, to capture only a specific aspect of economic complexity: for instance, when applied to countries, FC is mainly related to the degree of diversification of the export basket of each country, while MR essentially captures the similarities in the export baskets of the different countries (Sciarra, Chiarotti, Ridolfi, et al., 2020b).
The GENEPY index arises from the first two (normalized) eigenvectors (with the eigenvalues ordered in a weakly decreasing way) of a suitable symmetric proximity matrix, which is derived from an incidence matrix $\mathbf{M} \in \mathbb{R}^{C \times P}$ obtained by thresholding and binarizing the matrix of Revealed Comparative Advantage (RCA) values.

The two eigenvectors capture, respectively, information obtained by the FC method and the MR one.

The GENEPY index for countries is obtained in the following way (a similar construction holds for the GENEPY index for products).

i) First, for a specific year, the matrix $\mathbf{RCA} \in \mathbb{R}^{C\times P}$ of RCA values in that year is determined (see the Introduction for details).

ii) In order to extract topological information from the **RCA** matrix, an incidence matrix $\mathbf{M} \in \mathbb{R}^{C\times P}$ is generated, whose entries are defined as follows:

$$M_{c,p} := \begin{cases} 1\,, & \text{if } RCA_{c,p} \geq 1\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{A.60}$$

Then, its weighted version $\mathbf{W} \in \mathbb{R}^{C\times P}$ is considered, whose generic element is defined as $W_{c,p} := \frac{M_{c,p}}{k_c k'_p}$, where $k_c := \sum_{p=1}^{P} M_{c,p}$ is the degree of the country c in the graph represented by the incidence matrix $\mathbf{M}$, and $k'_p := \sum_{c=1}^{C} \frac{M_{c,p}}{k_c}$ represents the degree of the product p corrected by how easily that product is found within the subnetwork of countries.

iii) The matrix $\mathbf{N} \in \mathbb{R}^{\mathbf{C}\times\mathbf{C}}$ is constructed, whose elements N_{c,c^*} are defined as follows:

$$N_{c,c^*} := \begin{cases} \sum_{p=1}^{P} W_{c,p} W_{c^*,p}\,, & \text{if } c \neq c\text{,} \\ 0, & \text{otherwise}\,. \end{cases} \tag{A.61}$$

Due to the weighting involved in the construction of the matrix $\mathbf{W}$, the resulting matrix $\mathbf{N}$ is symmetric. Each entry N_{c,c^*} of $\mathbf{N}$ represents the proximity of the two corresponding countries c and c^*.

iv) The (normalized) eigenvectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^C$ associated with the two largest eigenvalues $\lambda_1 \geq \lambda_2 \geq 0$ of $\mathbf{N}$ are determined. Their components are denoted as $x_{c,1}$ and $x_{c,2}$, respectively, for $c = 1, \ldots, C$.

v) Then, the GENEPY index of country c for the specific year is defined as follows:

$$GENEPY_c := \left(\sum_{i=1}^{2} \lambda_i x_{c,i}^2 \right)^2 + 2 \sum_{i=1}^{2} \lambda_i^2 x_{c,i}^2\,. \tag{A.62}$$

The specific nonlinear transformation from $x_{c,1}$ and $x_{c,2}$ to $GENEPY_c$, which is used in Eq. (A.62), can be justified by rigorous statistical arguments, based on the use of the two (normalized) eigenvectors $\mathbf{x}_1$ and $\mathbf{x}_2$ to get a nonlinear least-square estimate of the matrix $\mathbf{N}$, and on the evaluation of how relevant $x_{c,i}$ and $x_{c,2}$ are

to obtain that estimate (Sciarra, Chiarotti, Laio, et al., 2018; Sciarra, Chiarotti, Ridolfi, et al., 2020b).

It is worth mentioning the qualitative difference between the GENEPY index and the ones determined by the FC and MR methods, considering again the case in which they are all applied to countries.

- The GENEPY index is highly related to a linearized version of the F index computed by the FC method(Sciarra, Chiarotti, Ridolfi, et al., 2020b), in which one searches for the (normalized) eigenvector associated with the largest eigenvalue of a slightly different matrix $\mathbf{N}_F \in \mathbb{R}^{C \times C}$ than the matrix $\mathbf{N}$. The specific matrix $\mathbf{N}_F$ is written as $\mathbf{N}_F := \mathbf{W}\mathbf{W}^\top$, where $\mathbf{W}$ is the same weighted incidence matrix considered in the context of the GENEPY index. The difference with respect to the case of the matrix $\mathbf{N}$ defined in Eq. (A.61) is that its diagonal entries are not set to 0 (in Eq. (A.61), such a choice of the diagonal entries is done in order to make the resulting $\mathbf{N}$ be a proximity matrix).
- The ECI index, computed by MR, is based on searching for the (normalized) eigenvector associated with the second-largest eigenvalue of a slightly different matrix $\mathbf{N}_{ECI} \in \mathbb{R}^{C \times C}$ than the matrix $\mathbf{N}$ considered by GENEPY. The specific matrix $\mathbf{N}_{ECI}$ is written as $\mathbf{N}_{ECI} := \mathbf{W}_{ECI}\mathbf{W}_{ECI}^\top$, where the elements of $\mathbf{W}_{ECI} \in \mathbb{R}^{C \times P}$ are defined as $W_{ECI,c,p} := \frac{M_{c,p}}{k_c k_p}$, being $k_p := \sum_{c=1}^{C} M_{c,p}$ the degree of the product p in the graph represented by the incidence matrix $\mathbf{M}$. The second-largest eigenvalue of $\mathbf{N}_{ECI}$ is considered, instead of its first-largest one, as one can show that the (normalized) eigevector associated with the latter is non-informative, for the specific matrix $\mathbf{N}_{ECI}$.

Similar comments hold for the case of the FC and MR methods when they are applied to products (obtaining, respectively, the Q index and the PCI index).

Bibliography

(N.d.). https://ungerconsulting.net/fda-enforcement-statistics-fy-2009-2015/. Accessed: 23.03.2019.

(N.d.). https://www.efmc.info/medchemwatch-2012-1/sme-2.php#:~:text=Drug%20Discovery%20Scientist-,Drug%20Discovery,now%20estimated%20at%20%241.8%20billion.. Accessed: 3.12.2020.

Acemoglu, Daron and Joshua Linn (2004). "Market size in innovation: theory and evidence from the pharmaceutical industry". In: *The Quarterly journal of economics* 119.3, pp. 1049–1090.

Aitken, Alexander C (1936). "IV.—On least squares and linear combination of observations". In: *Proceedings of the Royal Society of Edinburgh* 55, pp. 42–48.

al., M.Provost et (2019). "Pharmaceutical antitrust law in European Union." In: Dechert LLP.

Alfakih, Abdo and Henry Wolkowicz (2000). "Matrix completion problems". In: *Handbook of semidefinite programming*. Springer, pp. 533–545.

Almeida, Fernando LF (2017). "Benefits, challenges and tools of big data management". In: *Journal of Systems Integration* 8.4, pp. 12–20.

Alsharkas, Zeina (2014). "Firm size, competition, financing and innovation". In: *International Journal of Management and Economics* 44.1, pp. 51–73.

Andreß, Hans-Jürgen, Katrin Golsch, and Alexander W Schmidt (2013). *Applied panel data analysis for economic and social surveys*. Springer Science & Business Media.

Andrew, Crane-Droesch (2017). *Semiparametric Panel Data Using Neural Networks*. Tech. rep. Agricultural and Applied Economics Association.

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly harmless econometrics*. Princeton university press.

Arellano, Manuel (2003). *Panel data econometrics*. Oxford university press.

Athey, Susan, Mohsen Bayati, et al. (2021). "Matrix completion methods for causal panel data models". In: *Journal of the American Statistical Association*, pp. 1–15.

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.

– (2019). "Machine learning methods economists should know about, arxiv". In:

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Bai, Zhidong, Philip E Cheng, and Cun-Hui Zhang (1997). "An extension of the Hardy-Littlewood strong law". In: *Statistica Sinica*, pp. 923–928.

Bala, Ram, Pradeep Bhardwaj, and Pradeep K Chintagunta (2017). "Pharmaceutical product recalls: Category effects and competitor response". In: *Marketing Science* 36.6, pp. 931–943.

Balasubramanian, Natarajan and Jeongsik Lee (2008). "Firm age and innovation". In: *Industrial and Corporate Change* 17.5, pp. 1019–1047.

Ball, George, Jeffrey Thomas Macher, and Ariel Dora Stern (2018). "Recalls, innovation, and competitor response: Evidence from medical device firms". In:

Ball, George P, Rachna Shah, and Kaitlin D Wowak (2018). "Product competition, managerial discretion, and manufacturing recalls in the US pharmaceutical industry". In: *Journal of Operations Management* 58, pp. 59–72.

Baltagi, Badi H (2006). "Estimating an economic model of crime using panel data from North Carolina". In: *Journal of Applied econometrics* 21.4, pp. 543–547.

Barata, João Carlos Alves and Mahir Saleh Hussein (2012). "The Moore–Penrose pseudoinverse: A tutorial review of the theory". In: *Brazilian Journal of Physics* 42.1-2, pp. 146–165.

Bargagli Stoffi, Falco Johannes (2020). "Essays on applied machine learning". In:

Bargagli-Stoffi, Falco J, Jan Niederreiter, and Massimo Riccaboni (2021). "Supervised learning for the prediction of firm dynamics". In: *Data Science for Economics and Finance*. Springer, Cham, pp. 19–41.

Barigozzi, Matteo, Giorgio Fagiolo, and Diego Garlaschelli (2010). "Multinetwork of international trade: A commodity-specific analysis". In: *Physical Review E* 81.4, p. 046104.

Barlow, Roger J (1993). *Statistics: a guide to the use of statistical methods in the physical sciences*. Vol. 29. John Wiley & Sons.

Baumann, Julian and Alexander S Kritikos (2016). "The link between R&D, innovation and productivity: Are micro firms different?" In: *Research Policy* 45.6, pp. 1263–1274.

Baumol, William J, Alan S Blinder, and Edward N Wolff (2003). *Downsizing in America: Reality, causes, and consequences*. Russell Sage Foundation.

Belloni, Alexandre, Victor Chernozhukov, and Lie Wang (2011). "Square-root lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4, pp. 791–806.

Bennett, James, Stan Lanning, et al. (2007). "The netflix prize". In: *Proceedings of KDD cup and workshop*. Vol. 2007. New York, NY, USA., p. 35.

Bertoni, John M et al. (2010). "Increased melanoma risk in Parkinson disease: a prospective clinicopathological study". In: *Archives of Neurology* 67.3, pp. 347–352.

Bhargava, Alok, Luisa Franzini, and Wiji Narendranathan (1982). "Serial correlation and the fixed effects model". In: *The Review of Economic Studies* 49.4, pp. 533–549.

Bhatia, R (1997). *Matrix Analysis Springer-Verlag New York*.

Blume-Kohout, Margaret E and Neeraj Sood (2013). "Market size and innovation: Effects of Medicare Part D on pharmaceutical research and development". In: *Journal of public economics* 97, pp. 327–336.

Boelaert, Julien and Étienne Ollion (2018). "The great regression". In: *Revue française de sociologie* 59.3, pp. 475–506.

Bowe C. (2005). *Merck quarterly profits hit by Vioxx recall*. [Online; accessed 15-April-2021]. URL: `https://www.ft.com/content/b507ea92-b25b-11d9-bcc6-00000e2511c8`.

Breiman, Leo (2001). "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3, pp. 199–231.

Bresnahan, Timothy F and Peter C Reiss (1991). "Entry and competition in concentrated markets". In: *Journal of political economy* 99.5, pp. 977–1009.

Burke, Mary A and Tim R Sass (2013). "Classroom peer effects and student achievement". In: *Journal of Labor Economics* 31.1, pp. 51–82.

Cai, Jian-Feng, Emmanuel J Candès, and Zuowei Shen (2010). "A singular value thresholding algorithm for matrix completion". In: *SIAM Journal on optimization* 20.4, pp. 1956–1982.

Cameron, A Colin and Pravin K Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

– (2013). *Regression analysis of count data*. Vol. 53. Cambridge university press.

Candès, Emmanuel J and Benjamin Recht (2009). "Exact matrix completion via convex optimization". In: *Foundations of Computational mathematics* 9.6, pp. 717–772.

Cerda, Rodrigo A (2007). "Endogenous innovations in the pharmaceutical industry". In: *Journal of Evolutionary Economics* 17.4, pp. 473–515.

Chamberlain, Gary (2000). "Econometrics and decision theory". In: *Journal of Econometrics* 95.2, pp. 255–283.

Chen, Chun-Hung and Loo Hay Lee (2011a). *Stochastic simulation optimization: an optimal computing budget allocation*. Vol. 1. World scientific.

– (2011b). *Stochastic simulation optimization: an optimal computing budget allocation*. Vol. 1. World scientific.

Cheng, Jessie (2008). "An antitrust analysis of product hopping in the pharmaceutical industry". In: *Colum. L. Rev.* 108, p. 1471.

Chernozhukov, Victor et al. (2017). "Double/debiased/neyman machine learning of treatment effects". In: *American Economic Review* 107.5, pp. 261–65.

Civan, Abdulkadir and Michael T Maloney (2009). "The effect of price on pharmaceutical R&D". In: *The BE Journal of Economic Analysis & Policy* 9.1.

Consoli, Sergio, Diego Reforgiato Recupero, and Michaela Saisana (2021). *Data Science for Economics and Finance: Methodologies and Applications*.

Dash, Sabyasachi et al. (2019). "Big data in healthcare: management, analysis and future prospects". In: *Journal of Big Data* 6.1, pp. 1–25.

Dattorro, Jon (2010). *Convex optimization & Euclidean distance geometry*. Lulu. com.

Davenport, Thomas H and Jill Dyché (2013). "Big data in big companies". In: *International Institute for Analytics* 3.1-31.

Davis, Jonathan and Sara B Heller (2017). "Using causal forests to predict treatment heterogeneity: An application to summer jobs". In: *American Economic Review* 107.5, pp. 546–50.

Dear, Carley et al. (2021). "Grand Challenges Initiative: Producing Actionable and Visible Results that Drive Innovation and Improvement". In: *Intersection: A Journal at the Intersection of Assessment and Learning* 2.3, p. 24576.

DiMasi, Joseph A, Ronald W Hansen, and Henry G Grabowski (2003). "The price of innovation: new estimates of drug development costs". In: *Journal of health economics* 22.2, pp. 151–185.

Dominici, Francesca, Falco J Bargagli-Stoffi, and Fabrizia Mealli (2020). "From controlled to undisciplined data: estimating causal effects in the era of data science using a potential outcome framework". In: *arXiv preprint arXiv:2012.06865*.

Dubé, Jean-Pierre and Sanjog Misra (2017). *Scalable price targeting*. Tech. rep. National Bureau of Economic Research.

Dubois, Pierre et al. (2015). "Market size and pharmaceutical innovation". In: *The RAND Journal of Economics* 46.4, pp. 844–871.

Duggan, Mark and Fiona Scott Morton (2010). "The effect of Medicare Part D on pharmaceutical prices and utilization". In: *American Economic Review* 100.1, pp. 590–607.

Efron, Bradley and Trevor Hastie (2016). *Computer age statistical inference*. Vol. 5. Cambridge University Press.

F., Nutarelli (2021). "At the edge of economics and Machine-Learning ((Unpublished doctoral dissertation)". In: *IMT Lucca for Advanced Studies*.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

FDA U.S. Food Drug (2019). *CFR - Code of Federal Regulations Title 21 (Subpart "A-General Provisions", Sec. 7.3 "Definitions") no. 21CFR7.3*. `https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=7.3`.

Fisher, Susan Reynolds and Margaret A White (2000). "Downsizing in a learning organization: are there hidden costs?" In: *Academy of Management Review* 25.1, pp. 244–251.

Florescu, Ionut (2014). *Probability and stochastic processes*. John Wiley & Sons.

Foucart, Simon et al. (2017). "De-biasing low-rank projection for matrix completion". In: *Wavelets and Sparsity XVII*. Vol. 10394. International Society for Optics and Photonics, p. 1039417.

Frees, Edward W et al. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press.

Gallelli, Luca et al. (2013). "Safety and efficacy of generic drugs with respect to brand formulation". In: *Journal of pharmacology & pharmacotherapeutics* 4.Suppl1, S110.

Geroski, Paul A and Chris F Walters (1995). "Innovative activity over the business cycle". In: *The Economic Journal* 105.431, pp. 916–928.

Gershgorin, Semyon Aranovich (1931). "Uber die abgrenzung der eigenwerte einer matrix". In: . 6, pp. 749–754.

Giaccotto, Carmelo, Rexford E Santerre, and John A Vernon (2005). "Drug prices and research and development investment behavior in the pharmaceutical industry". In: *The Journal of Law and Economics* 48.1, pp. 195–214.

Gilgen, Hans (2006). *Univariate time series in geosciences: theory and examples*. Springer Science & Business Media.

Gnecco, Giorgio, Stefano Amato, et al. (2020). "Machine Learning Application to Family Business Status Classification". In: *International Conference on Machine Learning, Optimization, and Data Science*. Springer, pp. 25–36.

Gnecco, Giorgio and Federico Nutarelli (2019a). "On the trade-off between number of examples and precision of supervision in machine learning problems". In: *Optimization Letters*, pp. 1–23.

– (2019b). "On the trade-off between number of examples and precision of supervision in regression problems: Proceedings of the INNS Big Data and Deep Learning Conference INNSBDDL2019, Held at Sestri Levante, Genova, Italy 16-18 April 2019". In:

– (2019c). “Optimal trade-off between sample size and precision of supervision for the fixed effects panel data model”. In: *International Conference on Machine Learning, Optimization, and Data Science*. Springer, pp. 531–542.

Gnecco, Giorgio, Federico Nutarelli, and Daniela Selvi (2020). “Optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model”. In: *Soft Computing* 24.21, pp. 15937–15949.

– (2021). “Optimal data collection design in machine learning: the case of the fixed effects generalized least squares panel data model”. In: *Machine Learning*, pp. 1–36.

Gray, Robert M (2006). “Toeplitz and circulant matrices: A review”. In:

Greene, William H (2003). *Econometric analysis*. Pearson Education India.

Groves, RM et al. (2004a). “Sample design and sampling error”. In: *Survey Methodology. Hoboken, NJ: Wiley-Interscience*, pp. 93–135.

– (2004b). “Sample design and sampling error”. In: *Survey Methodology. Hoboken, NJ: Wiley-Interscience*, pp. 93–135.

Günther, Wendy Arianne et al. (2017). “Debating big data: A literature review on realizing value from big data”. In: *The Journal of Strategic Information Systems* 26.3, pp. 191–209.

Hall, Bronwyn H and Nathan Rosenberg (2010). *Handbook of the Economics of Innovation*. Vol. 1. Elsevier.

Hall, Kelsey et al. (2016). “Characteristics of FDA drug recalls: A 30-month analysis”. In: *American Journal of Health-System Pharmacy* 73.4, pp. 235–240.

Hansen, Christian and Damian Kozbur (2014). “Instrumental variables estimation with many weak instruments using regularized JIVE”. In: *Journal of Econometrics* 182.2, pp. 290–308.

Hartford, Jason et al. (2016). “Counterfactual prediction with deep instrumental variables networks”. In: *arXiv preprint arXiv:1612.09596*.

Hashi, Iraj and Nebojša Stojčić (2013). “The impact of innovation activities on firm performance using a multi-stage model: Evidence from the Community Innovation Survey 4”. In: *Research Policy* 42.2, pp. 353–366.

Hastie, Tibshirani and Robert Tibshirani (2009). *& Friedman, J.(2008). The Elements of Statistical Learning; Data Mining, Inference and Prediction*.

Hastie, Trevor, Rahul Mazumder, et al. (2015). "Matrix completion and low-rank SVD via fast alternating least squares". In: *The Journal of Machine Learning Research* 16.1, pp. 3367–3402.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hellwig, Martin F and Klaus M Schmidt (2002). "Discrete–time approximations of the holmström–milgrom brownian–motion model of intertemporal incentive provision". In: *Econometrica* 70.6, pp. 2225–2264.

Herriges, Joseph (2011). *Micro-Econometrics: Methods of Moments and Limited Dependent Variables*.

Hidalgo, César A (2021). "Economic complexity theory and applications". In: *Nature Reviews Physics* 3.2, pp. 92–113.

Hidalgo, César A and Ricardo Hausmann (2009). "The building blocks of economic complexity". In: *Proceedings of the national academy of sciences* 106.26, pp. 10570–10575.

Hoover, Kevin D (2008). "Causality in economics and econometrics". In: *The new Palgrave dictionary of economics* 2.

Huergo, Elena and Jordi Jaumandreu (2004). "How does probability of innovation change with firm age?" In: *Small Business Economics* 22.3-4, pp. 193–207.

Im, Kyung So et al. (1999). "Efficient estimation of panel data models with strictly exogenous explanatory variables". In: *Journal of Econometrics* 93.1, pp. 177–201.

Imai, Kosuke and Aaron Strauss (2011). "Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign". In: *Political Analysis* 19.1, pp. 1–19.

Jacob, Schmookler (1966). "Invention and economic growth". In: *J. Schmookler.–Cambridge Mass: Harvard University Press*.

Jovanovic, Boyan (2020). *Product recalls and firm reputation*. Tech. rep. National Bureau of Economic Research.

Keeley, Lawrence H (1988). "Hunter-gatherer economic complexity and "population pressure": A cross-cultural analysis". In: *Journal of anthropological archaeology* 7.4, pp. 373–411.

Kiefer, Nicholas M (1980). "Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance". In: *Journal of econometrics* 14.2, pp. 195–202.

Kleinberg, Jon, Himabindu Lakkaraju, et al. (2018). "Human decisions and machine predictions". In: *The quarterly journal of economics* 133.1, pp. 237–293.

Kleinberg, Jon, Jens Ludwig, et al. (2015). "Prediction policy problems". In: *American Economic Review* 105.5, pp. 491–95.

Kleinknecht, Alfred and Bart Verspagen (1990). "Demand and innovation: Schmookler re-examined". In: *Research policy* 19.4, pp. 387–394.

Klepper, Steven and Franco Malerba (2010). "Demand, innovation and industrial dynamics: an introduction". In: *Industrial and Corporate Change* 19.5, pp. 1515–1520.

Kolluru, Srinivas and Pundarik Mukhopadhaya (2017). "Empirical studies on innovation performance in the manufacturing and service sectors since 1995: A systematic review". In: *Economic Papers: A journal of applied economics and policy* 36.2, pp. 223–248.

Kotthoff, Lars (2016). "Algorithm selection for combinatorial search problems: A survey". In: *Data Mining and Constraint Programming*. Springer, pp. 149–190.

Kyle, Margaret K and Anita M McGahan (2012). "Investments in pharmaceuticals before and after TRIPS". In: *Review of Economics and Statistics* 94.4, pp. 1157–1172.

Lanjouw, Jean O (2005). *Patents, price controls, and access to new drugs: how policy affects global market entry*. Tech. rep. National Bureau of Economic Research.

Lee, Brian K, Justin Lessler, and Elizabeth A Stuart (2010). "Improving propensity score weighting using machine learning". In: *Statistics in medicine* 29.3, pp. 337–346.

Li, Cai and Hua Zhou (2017). "svt: Singular value thresholding in MATLAB". In: *Journal of statistical software* 81.2.

Lichtenberg, Frank R (2006). "Pharmaceutical innovation as a process of creative destruction". In: *Knowledge Accumulation and Industry Evolution: The Case of Pharma-Biotech*, p. 61.

Lin, Wei and Jeffrey M Wooldridge (2019). "Testing and correcting for endogeneity in nonlinear unobserved effects models". In: *Panel Data Econometrics*. Elsevier, pp. 21–43.

Liu, Yan and Tian Xie (2019). "Machine learning versus econometrics: prediction of box office". In: *Applied Economics Letters* 26.2, pp. 124–130.

Luan, Chin-jung, Chengli Tien, and Yi-chuang Chi (2013). "Downsizing to the wrong size? A study of the impact of downsizing on firm performance during an economic downturn". In: *The International Journal of Human Resource Management* 24.7, pp. 1519–1535.

Lyko, Klaus, Marcus Nitzschke, and Axel-Cyrille Ngonga Ngomo (2016). "Big data acquisition". In: *New Horizons for a Data-Driven Economy*. Springer, Cham, pp. 39–61.

Ma, Wei and George H Chen (2019). "Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption". In: *arXiv preprint arXiv:1910.12774*.

Maciejewski, Anthony A and Charles A Klein (1985). "Obstacle avoidance for kinematically redundant manipulators in dynamically varying environments". In: *The international journal of robotics research* 4.3, pp. 109–117.

Malerba, Franco (2007). "Innovation and the evolution of industries". In: *Innovation, Industrial Dynamics and Structural Transformation*. Springer, pp. 7–27.

Markham, Anthony (2020). "Lurbinectedin: first approval". In: *Drugs*, pp. 1–9.

Martello, Silvano and Paolo Toth (1990). "Bin-packing problem". In: *Knapsack problems: Algorithms and computer implementations*, pp. 221–245.

Martin, Linda et al. (2017). *How much do clinical trials cost?*

Mas-Colell, Andreu, Michael Dennis Whinston, Jerry R Green, et al. (1995). *Microeconomic theory*. Vol. 1. Oxford university press New York.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani (2010a). "Spectral regularization algorithms for learning large incomplete matrices". In: *The Journal of Machine Learning Research* 11, pp. 2287–2322.

– (2010b). “Spectral regularization algorithms for learning large incomplete matrices”. In: *The Journal of Machine Learning Research* 11, pp. 2287–2322.

Mellahi, Kamel and Adrian Wilkinson (2010). “A study of the association between level of slack reduction following downsizing and innovation output”. In: *Journal of Management Studies* 47.3, pp. 483–508.

Mitchell, Tom M et al. (1997). “Machine learning”. In:

Mowery, David and Nathan Rosenberg (1979). “The influence of market demand upon innovation: a critical review of some recent empirical studies”. In: *Research policy* 8.2, pp. 102–153.

Mullainathan, Sendhil and Jann Spiess (2017). “Machine learning: an applied econometric approach”. In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Nguyen, Hung T et al. (2008a). “Trade-off between sample size and accuracy: case of measurements under interval uncertainty”. In:

– (2008b). “Trade-off between sample size and accuracy: case of measurements under interval uncertainty”. In:

NIH, National Institute on Aging (2020). *What Are Clinical Trials and Studies?* `https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies`.

Onakpoya, Igho J, Carl J Heneghan, and Jeffrey K Aronson (2016). “Worldwide withdrawal of medicinal products because of adverse drug reactions: a systematic review and analysis”. In: *Critical reviews in toxicology* 46.6, pp. 477–489.

Oneto, Luca et al. (2019). “Recent Advances in Big Data and Deep Learning: Proceedings of the INNS Big Data and Deep Learning Conference INNSBDDL2019, Held at Sestri Levante, Genova, Italy 16-18 April 2019”. In:

Pakes, Ariel and Mark Schankerman (1984). “The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources”. In: *R&D, patents, and productivity*. University of Chicago Press, pp. 73–88.

Pammolli, Fabio, Laura Magazzini, and Massimo Riccaboni (2011). “The productivity crisis in pharmaceutical R&D”. In: *Nature reviews Drug discovery* 10.6, pp. 428–438.

Parlett, Beresford N (1998). *The symmetric eigenvalue problem*. SIAM.

Pérez-Rodrıguez, Jorge V and Beatriz GL Valcarcel (2012). “Do product innovation and news about the R&D process produce large price changes and overreaction?

The case of pharmaceutical stock prices". In: *Applied Economics* 44.17, pp. 2217–2229.

Plonsky, Ori et al. (2017). "Psychological forest: Predicting human behavior". In: *Thirty-first AAAI conference on artificial intelligence*.

Price, W and II Nicholson (2014). "Making Do in Making Drugs: Innovation Policy and Pharmaceutical Manufacturing". In: *BCL Rev.* 55, p. 491.

Rake, Bastian (2017). "Determinants of pharmaceutical innovation: the role of technological opportunities revisited". In: *Journal of Evolutionary Economics* 27.4, pp. 691–727.

Rao, Calyampudi Radhakrishna et al. (1973). *Linear statistical inference and its applications*. Vol. 2. Wiley New York.

Révész, Pál (1968). *The laws of large numbers*. Tech. rep.

Roodman, David (2009). "How to do xtabond2: An introduction to difference and system GMM in Stata". In: *The stata journal* 9.1, pp. 86–136.

Ross, Paul F (1974). "Innovation Adoption by Organizations." In: *Personnel Psychology*.

Ruud, Paul A et al. (2000). "An introduction to classical econometric theory". In: *OUP Catalogue*.

Sanguineti, Marcello and Elena Tanfani (2021). *Special Issue Optimization for Machine Learning Guest Editorial*.

Scherer, Frederic M (1982). "Demand-pull and technological invention: Schmookler revisted". In: *The Journal of Industrial Economics*, pp. 225–237.

Schmookler, Jacob (2013). *Invention and economic growth*. Harvard University Press.

Sciarra, Carla, Guido Chiarotti, Francesco Laio, et al. (2018). "A change of perspective in network centrality". In: *Scientific reports* 8.1, pp. 1–9.

Sciarra, Carla, Guido Chiarotti, Luca Ridolfi, et al. (2020a). "Reconciling contrasting views on economic complexity". In: *Nature communications* 11.1, pp. 1–10.

– (2020b). "Reconciling contrasting views on economic complexity". In: *Nature communications* 11.1, pp. 1–10.

Sejnowski, Terrence J (2018). *The deep learning revolution*. MIT press.

Semykina, Anastasia and Jeffrey M Wooldridge (2010). "Estimating panel data models in the presence of endogeneity and selection". In: *Journal of Econometrics* 157.2, pp. 375–380.

Shalit, Uri, Fredrik D Johansson, and David Sontag (2017). "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Simoes, Alexander James Gaspar and César A Hidalgo (2011). "The economic complexity observatory: An analytical tool for understanding the dynamics of economic development". In: *Workshops at the twenty-fifth AAAI conference on artificial intelligence*.

Singh, Dilpreet and Chandan K Reddy (2015). "A survey on platforms for big data analytics". In: *Journal of big data* 2.1, pp. 1–20.

Siramshetty, Vishal B et al. (2016). "WITHDRAWN—a resource for withdrawn and discontinued drugs". In: *Nucleic acids research* 44.D1, pp. D1080–D1086.

Sivarajah, Uthayasankar et al. (2017). "Critical analysis of Big Data challenges and analytical methods". In: *Journal of Business Research* 70, pp. 263–286.

Stock, Gregory N, Noel P Greis, and William A Fischer (2002). "Firm size and dynamic technological innovation". In: *Technovation* 22.9, pp. 537–549.

Stoffi, Falco J Bargagli and Giorgio Gnecco (2018). "Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms". In: *Proceedings of the* 5^{th} *IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA 2018), Turin, Italy*, pp. 1–10.

– (2020). "Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms". In: *International Journal of Data Science and Analytics* 9.3, pp. 315–337.

Stoneman, Paul (2010). *Soft innovation: economics, product aesthetics, and the creative industries*. Oxford University Press.

Strang, Gilbert (1993). "The fundamental theorem of linear algebra". In: *The American Mathematical Monthly* 100.9, pp. 848–855.

Sun, Feng-Wen, Yimin Jiang, and John S Baras (2003). "On the convergence of the inverses of Toeplitz matrices and its applications". In: *IEEE Transactions on Information Theory* 49.1, pp. 180–190.

Symeonidis, George (1996). "Innovation, firm size and market structure: Schumpeterian hypotheses and some new themes". In:

Tacchella, Andrea et al. (2012). "A new metrics for countries' fitness and products' complexity". In: *Scientific reports* 2.1, pp. 1–7.

Terence N. (2004). *Merck Pulls Vioxx Painkiller From Market, and Stock Plunges*. [Online; accessed 15-April-2021]. URL: `https://www.nytimes.com/2004/09/30/business/merck-pulls-vioxx-painkiller-from-market-and-stock-plunges.html`.

Thach, Nguyen Ngoc, Vladik Kreinovich, and Nguyen Duc Trung (2021). *Data Science for Financial Econometrics*. Springer.

Thirumalai, Sriram and Kingshuk K. Sinha (2011). "Product Recalls in the Medical Device Industry: An Empirical Exploration of the Sources and Financial Consequences". In: *Management Science* 57, pp. 376–392.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tole, Alexandru Adrian et al. (2013). "Big data challenges". In: *Database systems journal* 4.3, pp. 31–40.

Tong, Carl H, Lee-Ing Tong, and James E Tong (2009). "The Vioxx recall case and comments". In: *Competitiveness Review: An International Business Journal*.

Vaishnav, Anish (2011). "Product Market Definition in Pharmaceutical Antitrust Cases: Evaluating Cross-Price Elasticity of Demand". In: *Colum. Bus. L. Rev.*, p. 586.

Vandenberghe, Lieven and Stephen Boyd (2004). *Convex optimization*. Vol. 1. Cambridge University Press Cambridge.

Vapnik, V (1998). "Statistical learning theory new york". In: *NY: Wiley* 1, p. 2.

Varian, Hal (2014). "Machine Learning and Econometrics". In: *Slides package from talk at University of Washington*.

Varian, Hal R (2014a). "Big data: New tricks for econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.

– (2014b). "Big data: New tricks for econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.

Vujović, Željko (2021). "A case study of the application of WEKA software to solve the problem of liver inflamation". In:

Wedin, Per-Åke (1973). "Perturbation theory for pseudo-inverses". In: *BIT Numerical Mathematics* 13.2, pp. 217–232.

Wikipedia contributors (2021). *Overfitting — Wikipedia, The Free Encyclopedia*. [Online; accessed 16-July-2021]. URL: `https://en.wikipedia.org/w/index.php?title=Overfitting&oldid=1016721642`.

William, H Greene (2008). "Econometric Analysis. 6". In:

Wooldridge, Jeffrey M (2002). "Econometric analysis of cross section and panel data MIT press". In: *Cambridge, MA* 108.

– (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wu, Xindong et al. (2008). "Top 10 algorithms in data mining". In: *Knowledge and information systems* 14.1, pp. 1–37.

Ying, Xue (2019). "An overview of overfitting and its solutions". In: *Journal of Physics: Conference Series*. Vol. 1168. 2. IOP Publishing, p. 022022.